

LINEAIRE REGRESSIE: DUMMIES EN INTERACTIES

30 mei 2024


Training O + S

Elmar Jansen (elmar@elmarjansen.nl)

VANDAAG

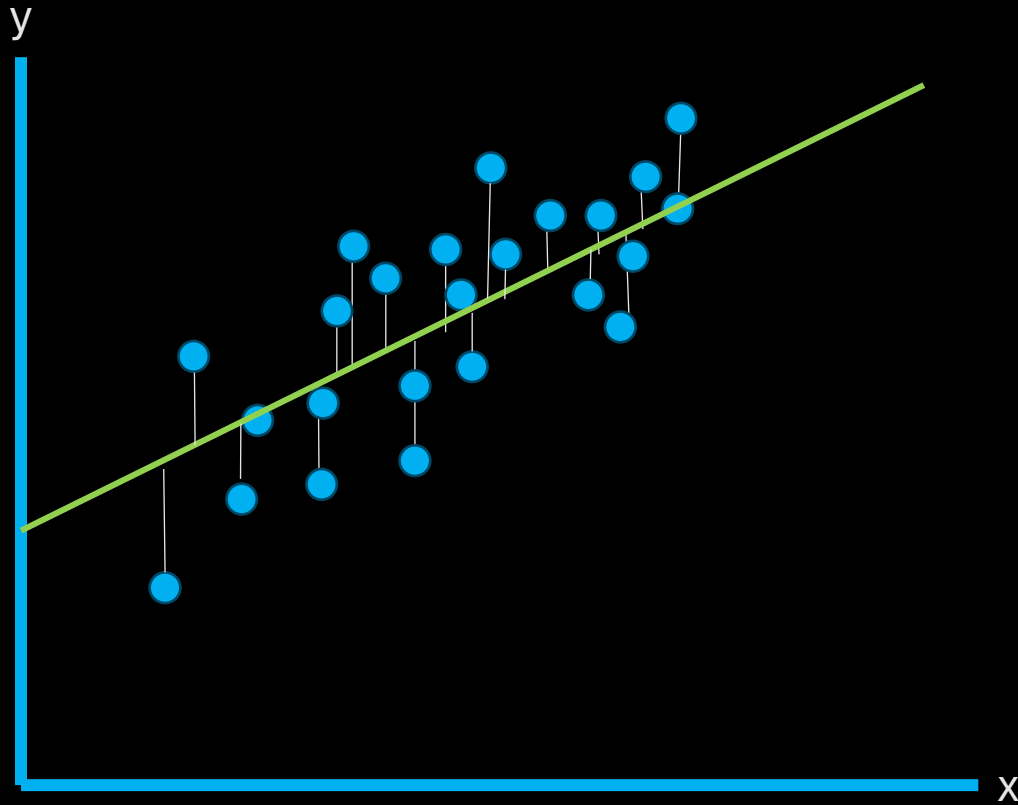
1. Terugblik
2. Dummies voor dichotome variabelen
3. Dummies voor nominale variabelen
4. Interactie met dummy
5. Interactie met continue variabele

DE KOMENDE WEKEN

Bijeenkomst	Onderwerp
Dinsdag 14 mei	Lineaire regressie: de basis
Dinsdag 21 mei	Lineaire regressie vervolg: assumpties en controleren
 Donderdag 30 mei	Interacties en dummy-variabelen
Dinsdag 4 juni	Logistische Regressie
Dinsdag 11 juni	Multilevel-analyse

TERUGBLIK VORIGE WEEK

LINEAIRE REGRESSIE



$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

INTERPRETATIE VAN COËFFICIËNTEN

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \varepsilon_i$$

Interpretatie van coëfficiënt β_0 :

Constante of intercept:

Waarde van afhankelijke variabele y als alle onafhankelijke variabelen x exact 0 zijn

Interpretatie van coëfficiënten β_1, β_2 etc.:

Effect of helling van variabele x_1 of x_2 , etc.

Stijging van y als deze onafhankelijke variabele stijgt met 1 (en overige onafhankelijke variabelen gelijk blijven)

OVERIGE BEGRIPPEN

Begrip	Beschrijving
R^2	Verklaarde variantie ("ons model voorspelt ..% van de variatie in afhankelijke variabele")
Adjusted R^2	R^2 gecorrigeerd voor aantal onafhankelijke variabelen en (kleine) steekproefomvang
Standaardfout	Standaardafwijking van schatting bij (hypothetisch) herhaald trekken van steekproeven (steekproevenverdeling). Interpretatie: verwachte gemiddelde afwijking van schatting.
T-toets	Statistische toets voor specifieke coëfficiënt. Nulhypothese: coëfficiënt is 0 (geen effect) in populatie.
F-toets	Statistische toets voor gehele model. Nulhypothese: model biedt geen enkele voorspellende waarde ($R^2=0$) in populatie.
Gestandaardiseerde Bèta	Coëfficiënt wanneer alle variabelen gestandaardiseerd zouden zijn. Interpretatie: aantal standaarddeviaties dat afhankelijke variabele stijgt als onafhankelijke variabele 1 standaarddeviatie stijgt. Ligt altijd tussen -1 en 1.

DANGER

8 GEVAREN VAN REGRESSIE

DANGER

DANGER!!

1



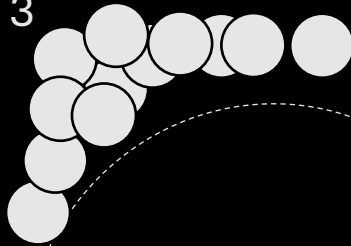
Schijnverband

2



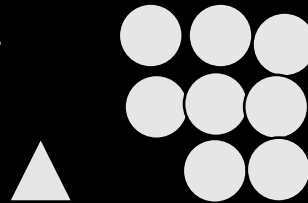
Wederkerigheid /
Simultaniteit

3



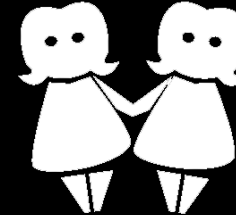
Non-Lineairiteit

4



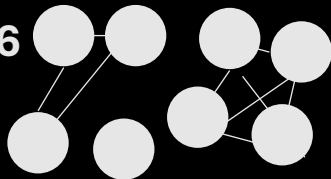
Extreme waarden

5



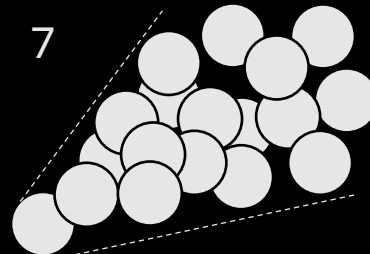
Multicollineariteit

6



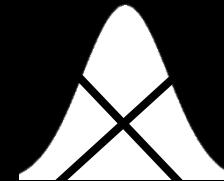
Niet onafhankelijke
residuen

7



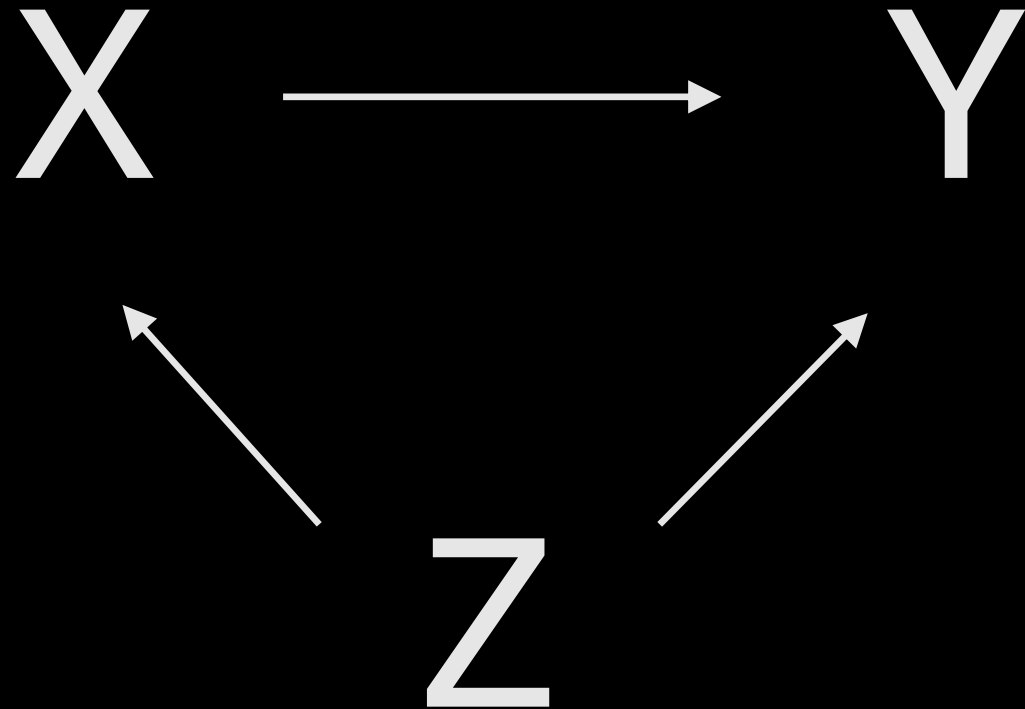
Heteroskedasticiteit

8



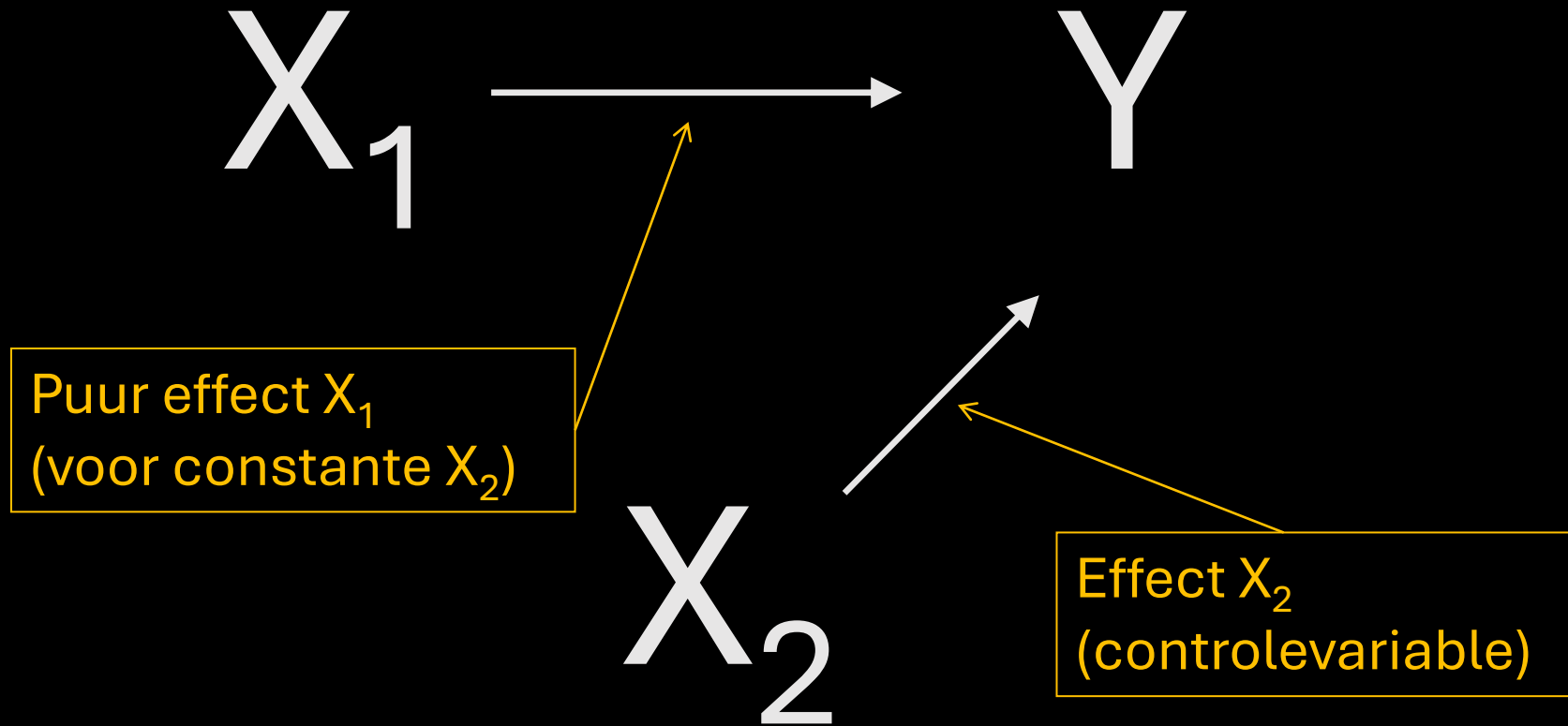
Non-normaliteit
van residuen

SCHIJNVERBAND



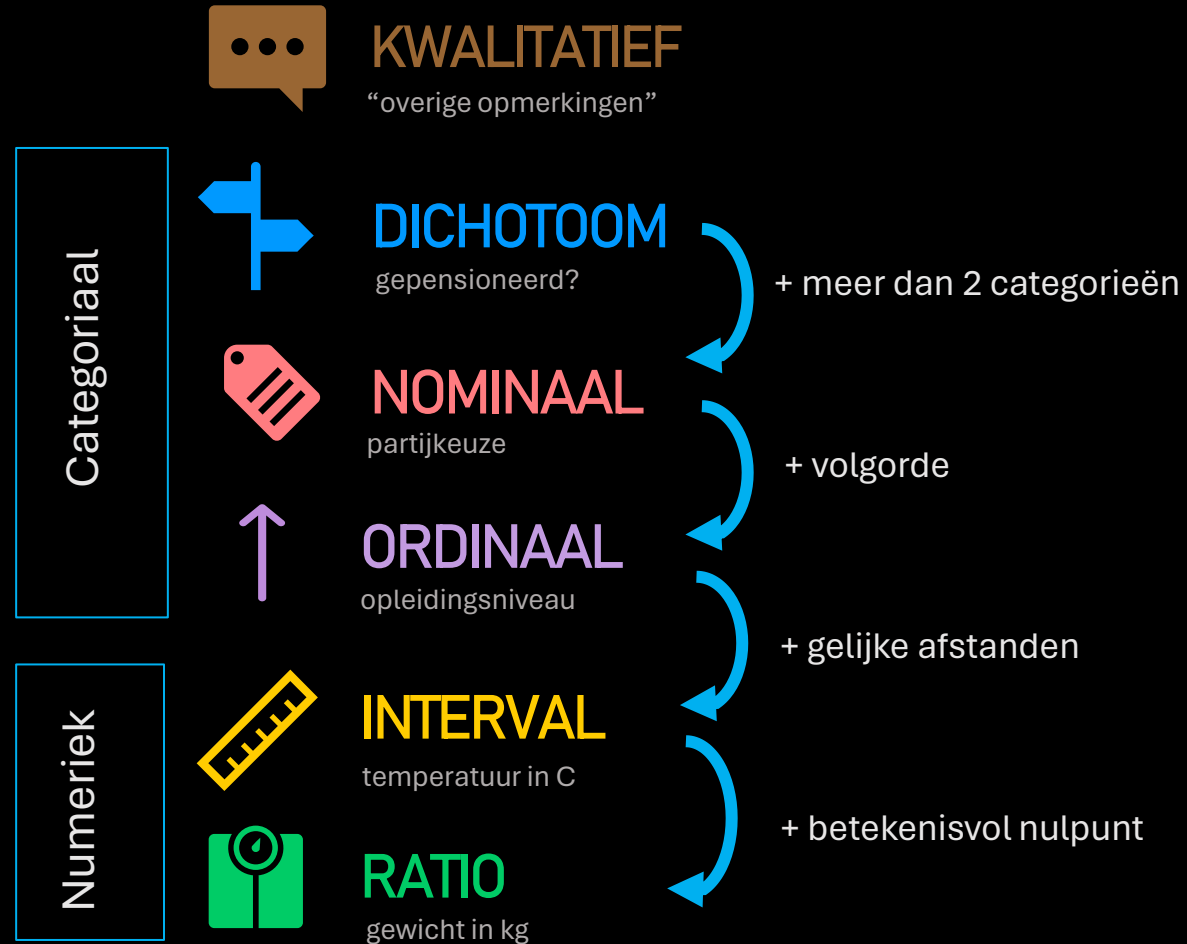
SCHIJNVERBAND: CONTROLEREN

Door de schijnverband-variabele “Z” toe te voegen aan het model kunnen we het probleem verhelpen:
we krijgen het effect **constant houdend voor Z**



DUMMY-VARIABLEN

HOE ZAT HET OOK ALWEER MET MEETNIVEAUS?



MEETNIVEAUS IN REGRESSIE

Dummy-variabelen zijn een trucje om categoriale variabelen als **onafhankelijke variabele** in regressie te gebruiken

In principe alleen numerieke variabelen als afhankelijke en onafhankelijke variabele

Categoriaal



KWALITATIEF

“overige opmerkingen”



DICHOTOOM

gepensioneerd?



NOMINAAL

partijkeuze



ORDINAAL

opleidingsniveau



INTERVAL

temperatuur in C



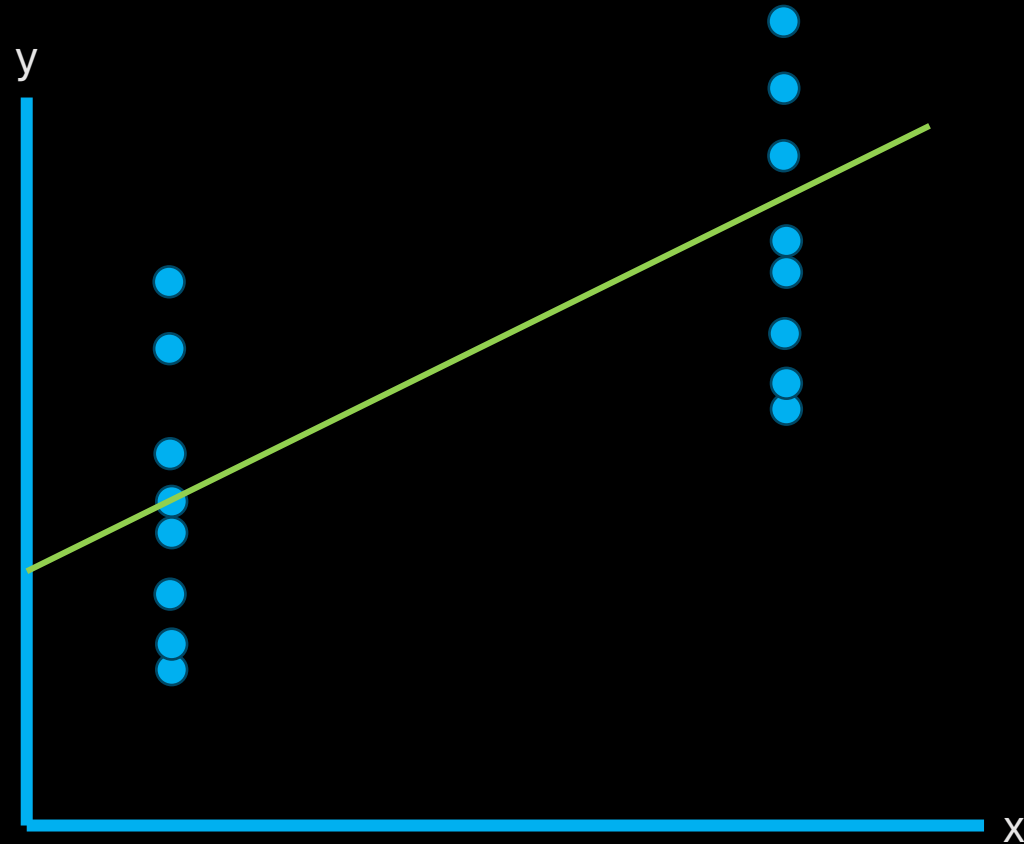
RATIO

gewicht in kg

Numeriek

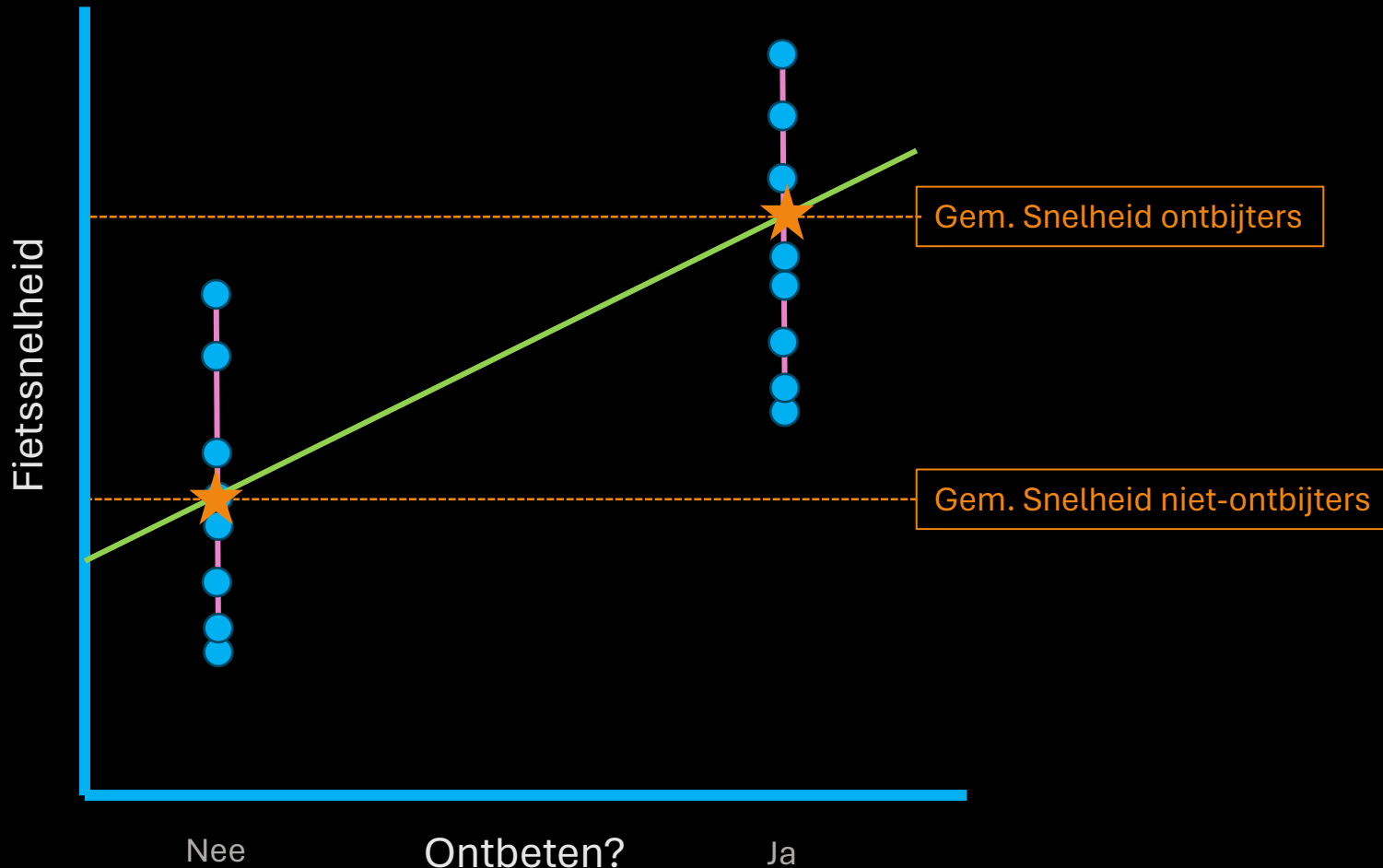
STAP 1: DICHOTOME VARIABLE ALS OV

Vraag: hoe ziet de regressie-grafiek eruit als we een **dichotome** onafhankelijke variabele zouden hebben?



STAP 1: DICHOTOME VARIABELE ALS OV

Voorbeeld: ontbijt en fietssnelheid



Waar kruist de lijn nu de punten?

➤ Op het *groepsgemiddelde*

Wat is nu het residu ε_i ?

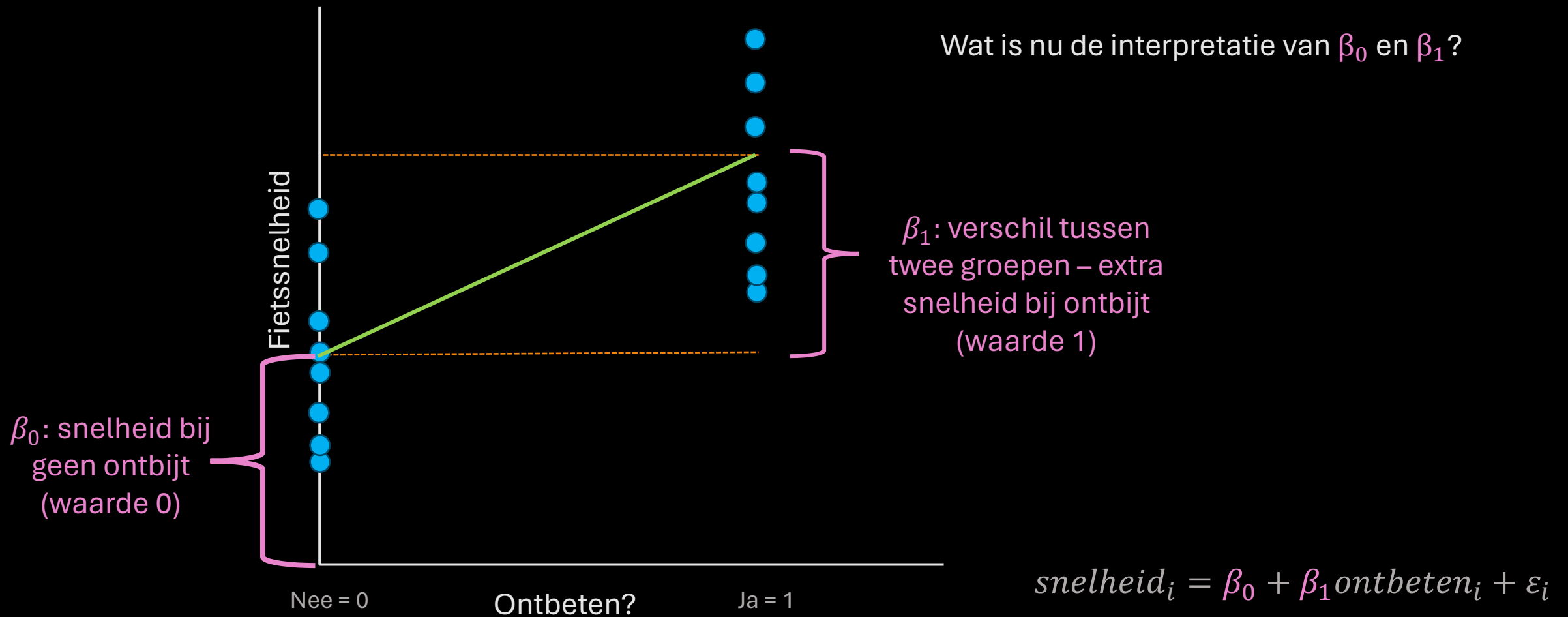
➤ De *afwijking* tot het groepsgemiddelde

Bij dichotome onafhankelijke variabele trekt regressie dus vanzelf een lijn tussen de twee groepsgemiddelde

$$snelheid_i = \beta_0 + \beta_1 ontbeten_i + \varepsilon_i$$

STAP 2: DE DUMMY-TRUC

We maken een dichotome variabele met waarden 0 en 1



STAP 2: DE DUMMY-TRUC

We maken een dichotome variabele met waarden 0 en 1

$$snelheid_i = \beta_0 + \beta_1 ontbeten_i + \varepsilon_i$$

Niet ontbeten: $snelheid_i = \beta_0 + \beta_1 \times 0 + \varepsilon_i = \boxed{\beta_0 + \varepsilon_i}$

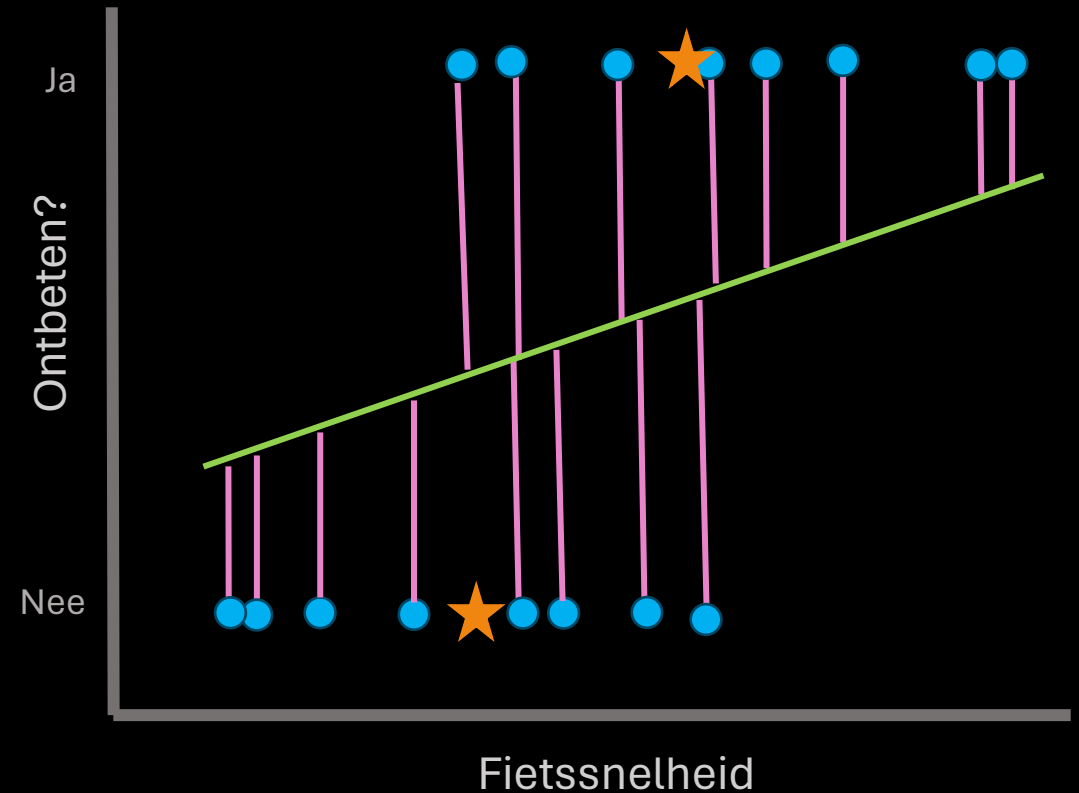
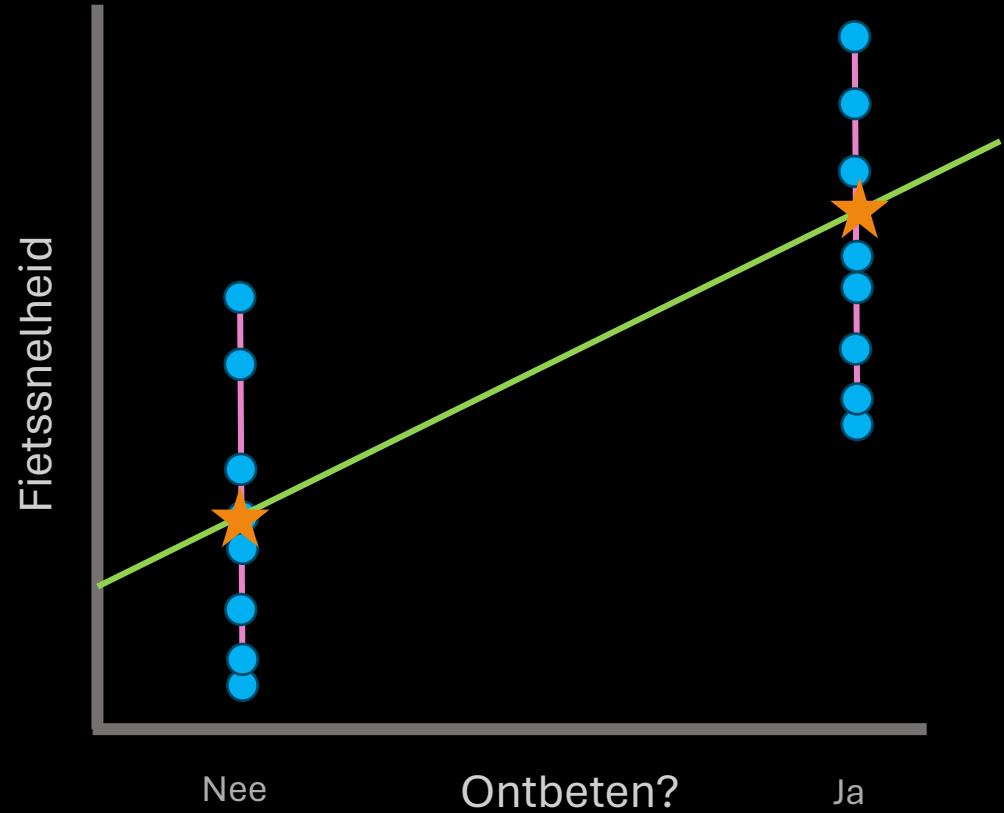
β_0 : snelheid bij
geen ontbijt
(waarde 0)

Wel ontbeten: $snelheid_i = \beta_0 + \beta_1 \times 1 + \varepsilon_i = \boxed{\beta_0 + \beta_1 + \varepsilon_i}$

β_1 : verschil tussen
twee groepen – extra
snelheid bij ontbijt
(waarde 1)

LET OP: ALLEEN VOOR ONAFHANKELIJKE VARIABLEN

We kunnen deze truc niet gebruiken voor afhankelijke variabelen!

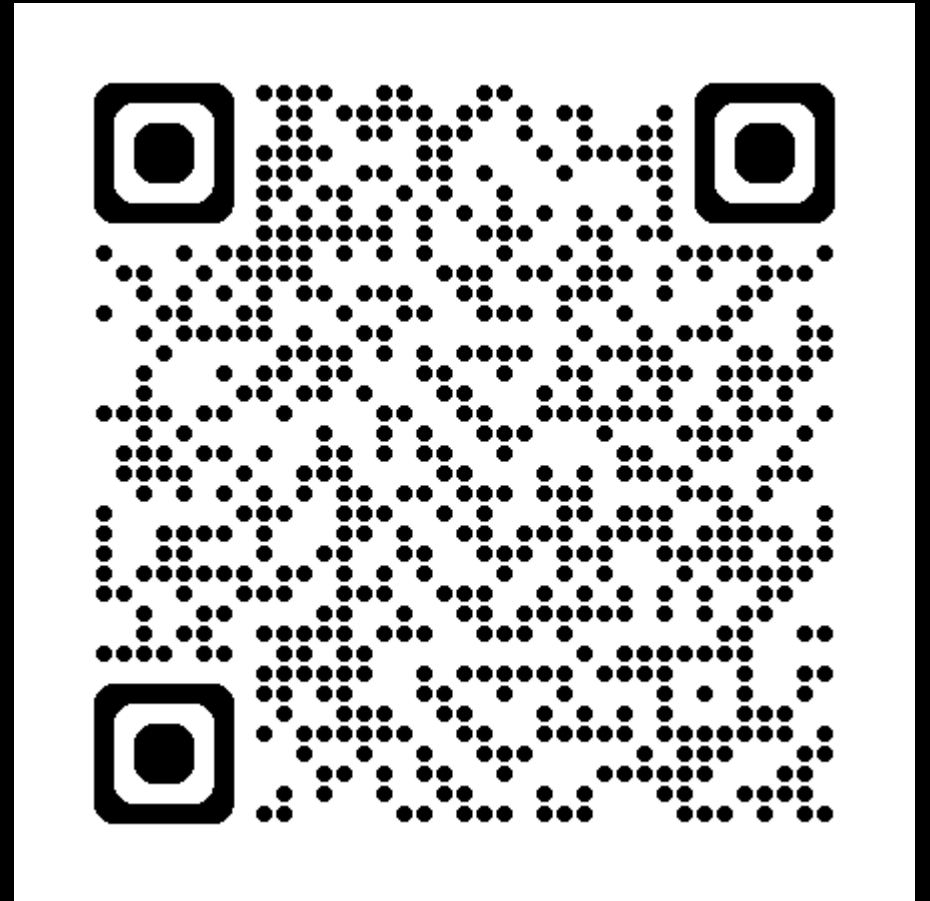


- Je krijgt dan een **andere (vlakkere) lijn** omdat we altijd de *verticale* residuen minimaliseren – hij loopt niet meer door de gemiddelden
- We voorspellen **onmogelijke uitkomsten** (y zal ook andere waarden aannemen dan alleen 0 en 1)
- Oplossing: **logistische regressie**

<https://elmarjansen.nl/os>

OEFENING 1

Geluk en werkloosheid



DUMMIES VOOR NOMINALE VARIABELEN

CATEGORIALE ONAFHANKELIJKE VARIABELEN

Corporatie-
huur

$$Koop_i = 0$$

$$PartHuur_i = 0$$

$$\text{CorpHuur}_i = 1$$

Particuliere
huur

$$Koop_i = 0$$

$$PartHuur_i = 1$$

$$\text{CorpHuur}_i = 0$$

Koop

$$Koop_i = 1$$

$$PartHuur_i = 0$$

$$\text{CorpHuur}_i = 0$$

(maakt onderscheid koop en huur)

(maakt onderscheid binnen huur)

(niet meer nodig)



Baseline-categorie

Je hebt altijd 1 dummy-variabele minder nodig dan er categorieën zijn. De categorie zonder eigen dummy wordt automatisch de “baseline”-categorie waar we de andere categorieën tegen afzetten.

MEERDERE DUMMIES

$$y_i = \beta_0 + \beta_1 \text{PartHuur}_i + \beta_2 \text{Koop}_i + \varepsilon_i$$

Corporatie-
huur

$$\begin{aligned}\text{PartHuur}_i &= 0 \\ \text{Koop}_i &= 0\end{aligned}$$

$$\begin{aligned}y_i &= \beta_0 + \beta_1 \times 0 + \beta_2 \times 0 + \varepsilon_i \\ y_i &= \beta_0 + \varepsilon_i\end{aligned}$$

Particuliere
huur

$$\begin{aligned}\text{PartHuur}_i &= 1 \\ \text{Koop}_i &= 0\end{aligned}$$

$$\begin{aligned}y_i &= \beta_0 + \beta_1 \times 1 + \beta_2 \times 0 + \varepsilon_i \\ y_i &= \beta_0 + \beta_1 + \varepsilon_i\end{aligned}$$

Koop

$$\begin{aligned}\text{PartHuur}_i &= 0 \\ \text{Koop}_i &= 1\end{aligned}$$

$$\begin{aligned}y_i &= \beta_0 + \beta_1 \times 0 + \beta_2 \times 1 + \varepsilon_i \\ y_i &= \beta_0 + \beta_2 + \varepsilon_i\end{aligned}$$

Interpretatie

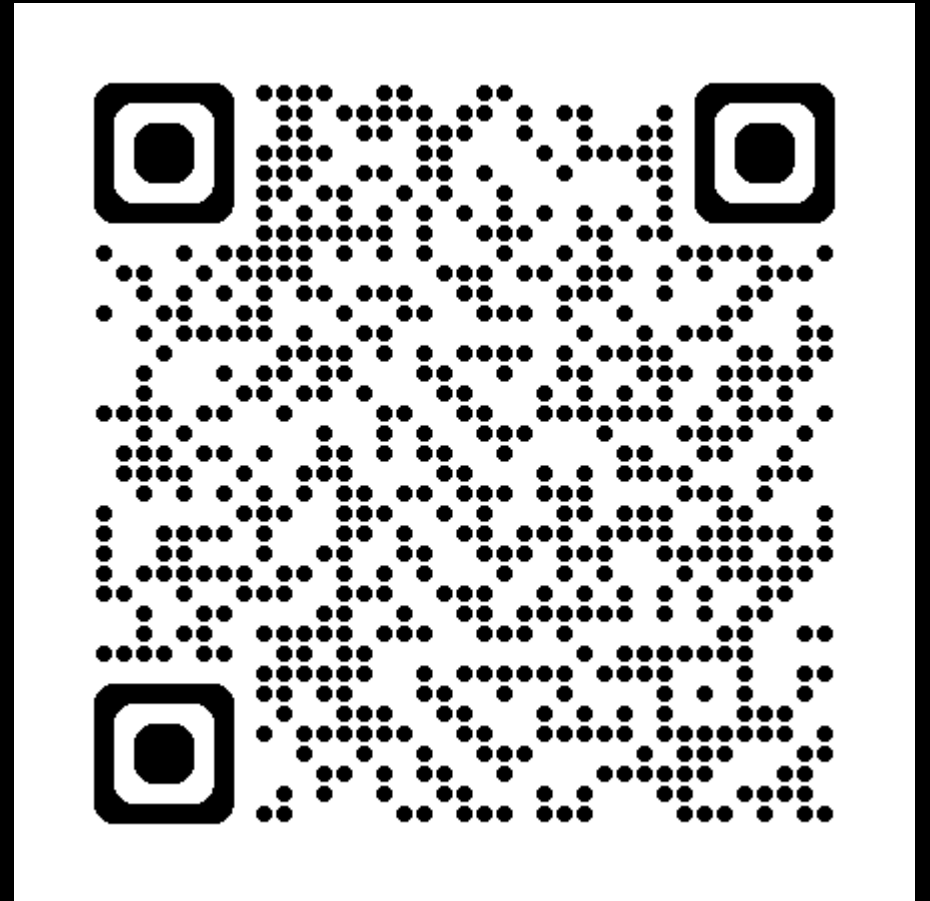
Coefficient van particuliere-huur-dummy (β_1):
verschil tussen corporatiehuur en particuliere huur

Coefficient van koop-dummy (β_2):
verschil tussen corporatiehuur en koop

<https://elmarjansen.nl/os>

OEFENING 2

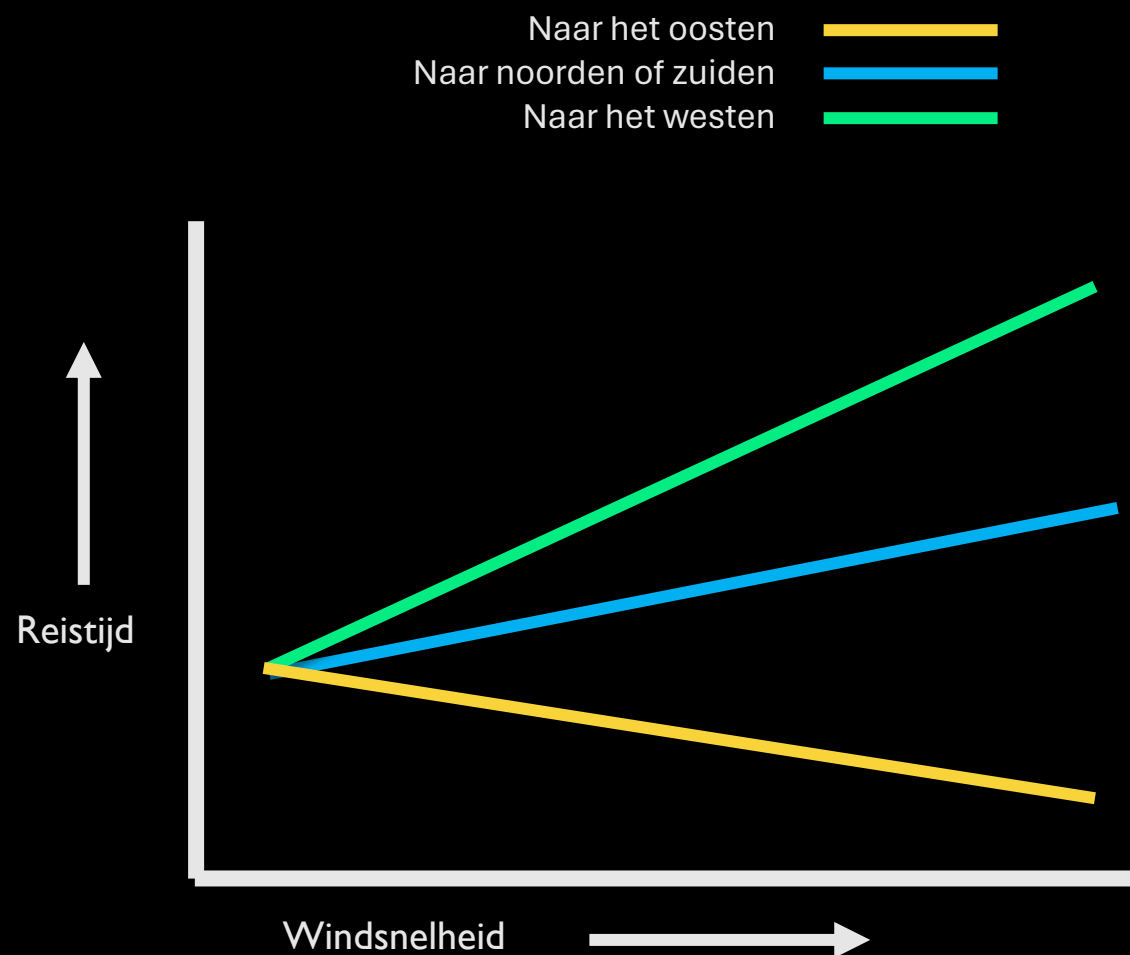
Geluk en woonplaats



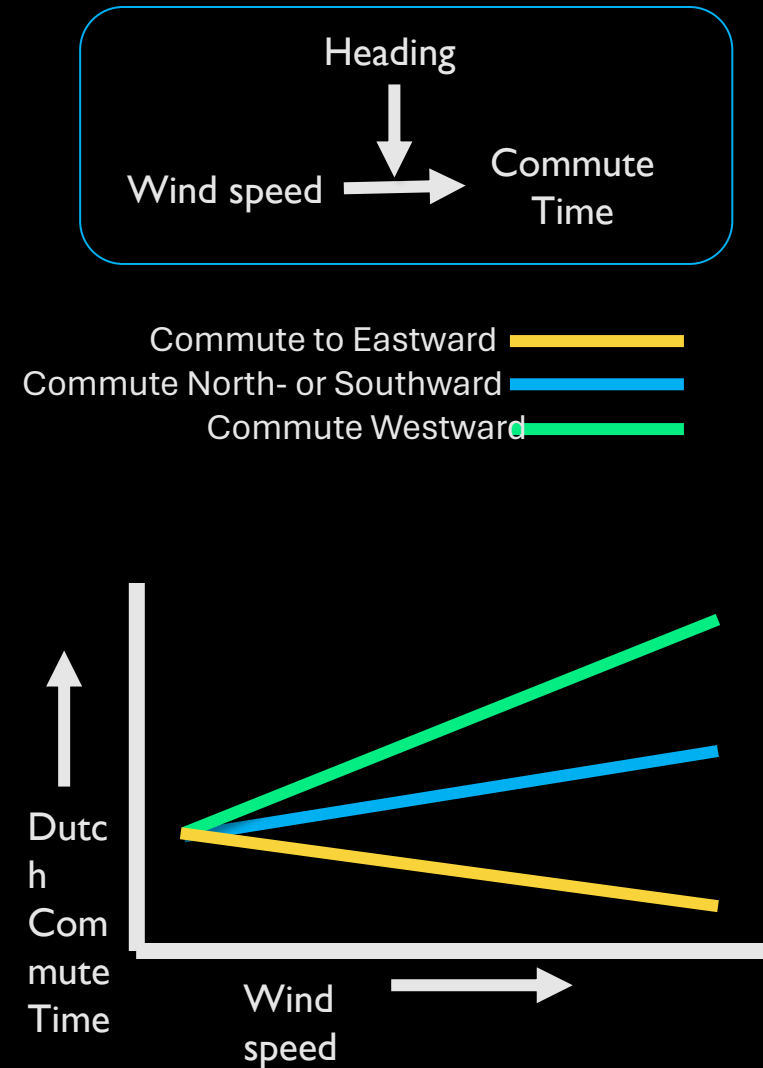
INTERACTIES

WAT IS EEN INTERACTIE- EFFECT?

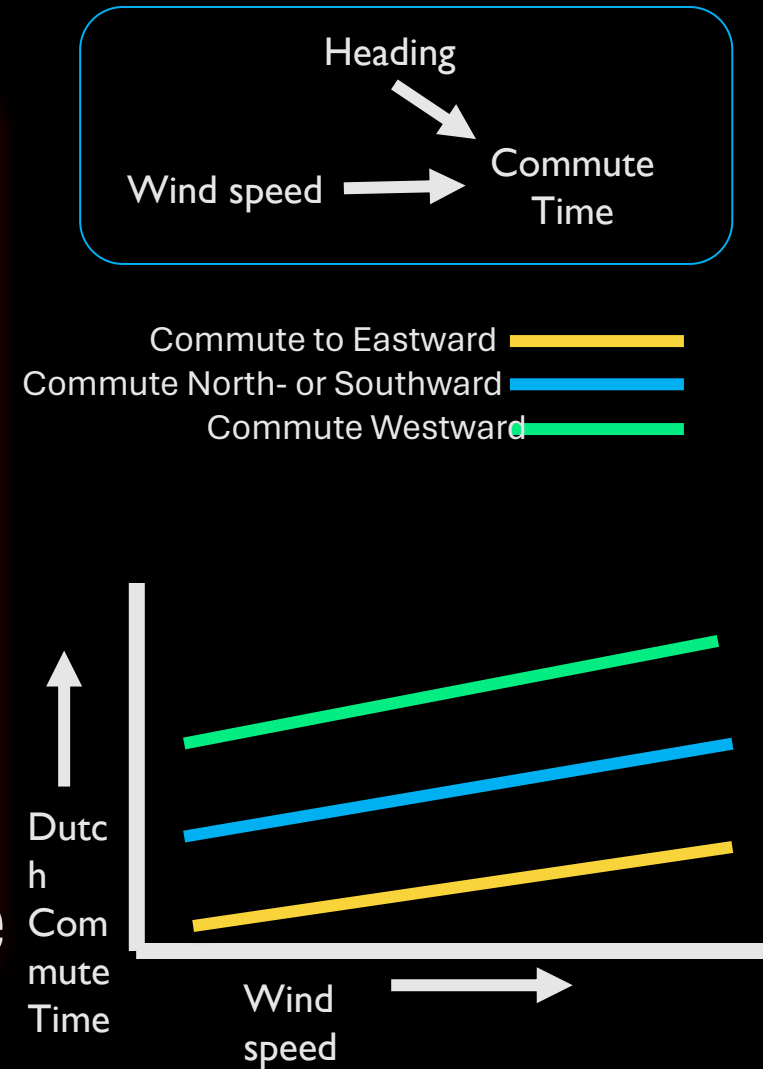




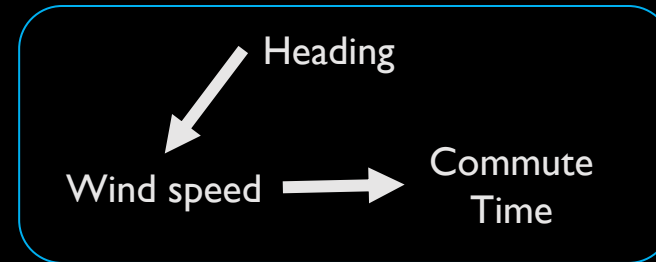
Interactie is:
Het *effect* van
een X varieert,
afhankelijk van
waarde van
andere X



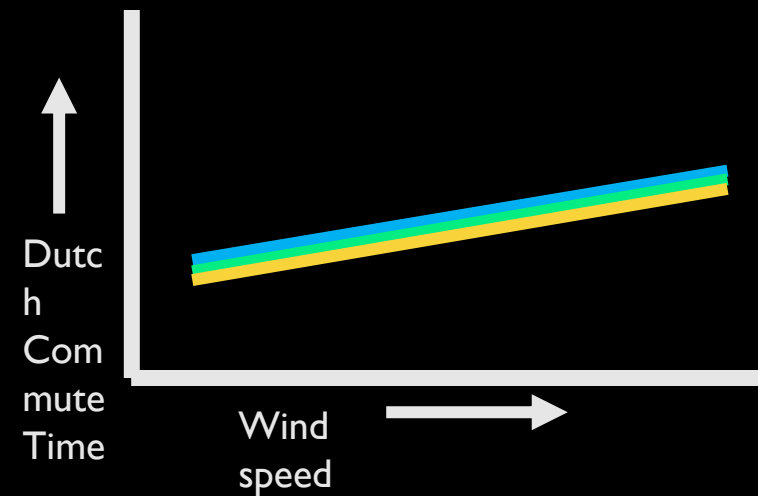
Interactie is niet
twee X-en
(onafhankelijke
variabelen)
die allebei een effect
hebben op de
afhankelijke variabele



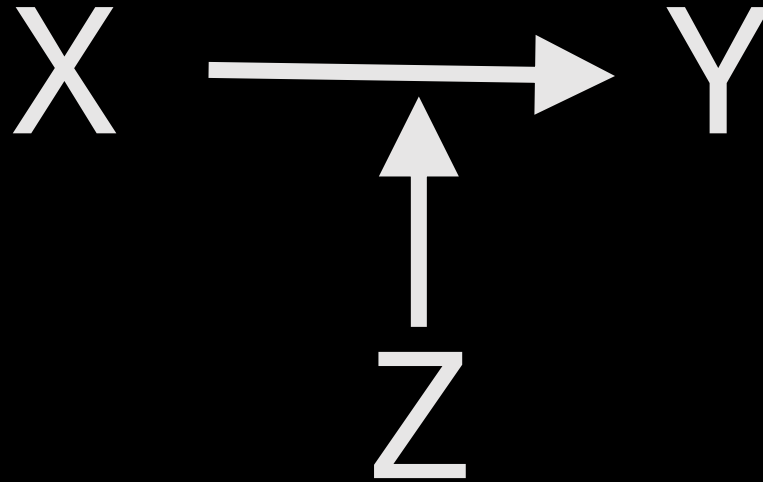
Interactie is niet
een X met effect
op een andere X



Commute to Eastward ———
Commute North- or Southward ———
Commute Westward ———



INTERACTIE IN SCHEMA



Een interactie-effect is een *effect op de omvang van een ander effect*.
Oftewel: hoe groot het effect van X op Y is, hangt af van Z.

Interactie

x_1 en x_2
interacteren

Contingentie

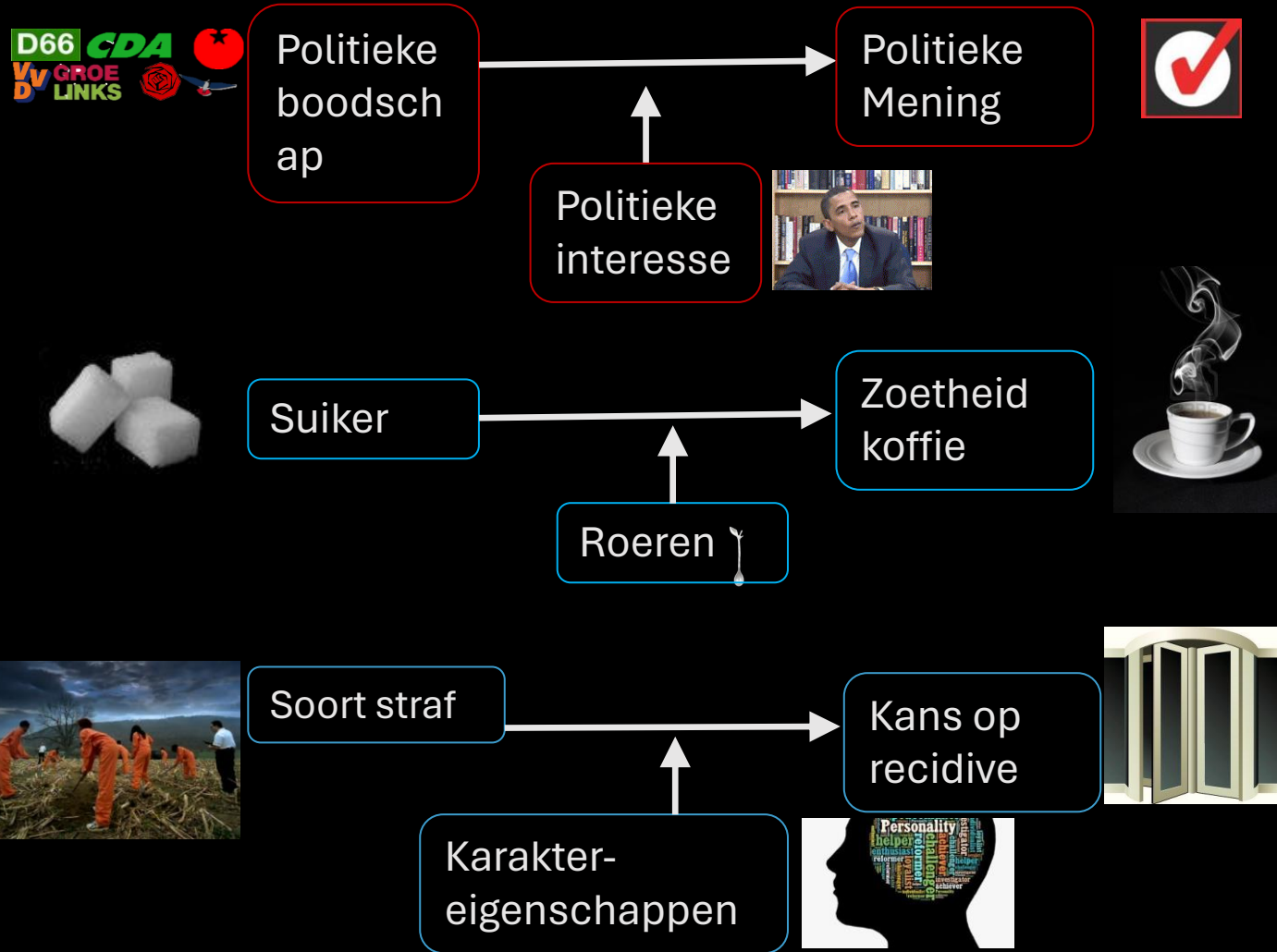
the effect of x_1 is contingent on
 x_2

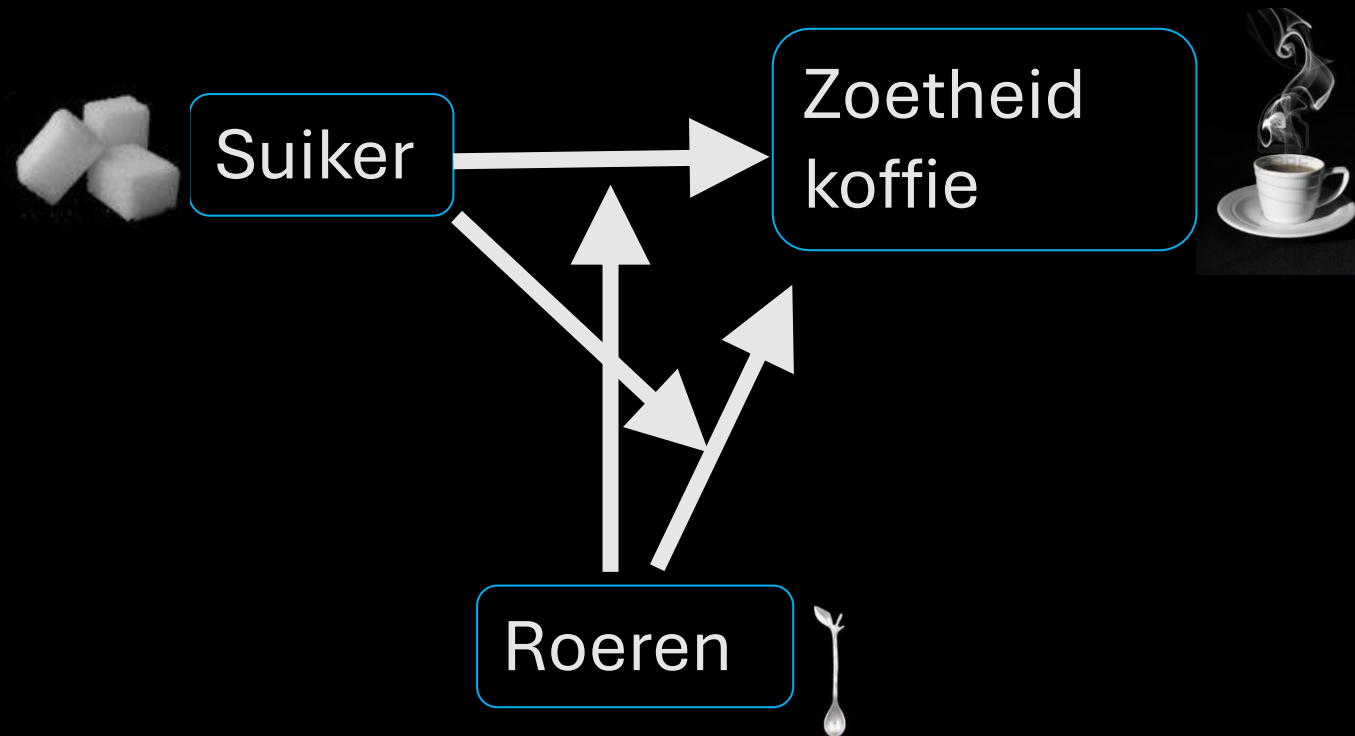
Moderatie

x_1 moderates the effect of x_2

Conditionaliteit

the effect of x_1 is conditional on x_2





Nerd-note: zeggen dat Z een interactie heeft op het effect van X op Y, is technisch hetzelfde als zeggen dat X een interactie heeft op het effect van Z op Y



INTERACTIE: VOORBEELD



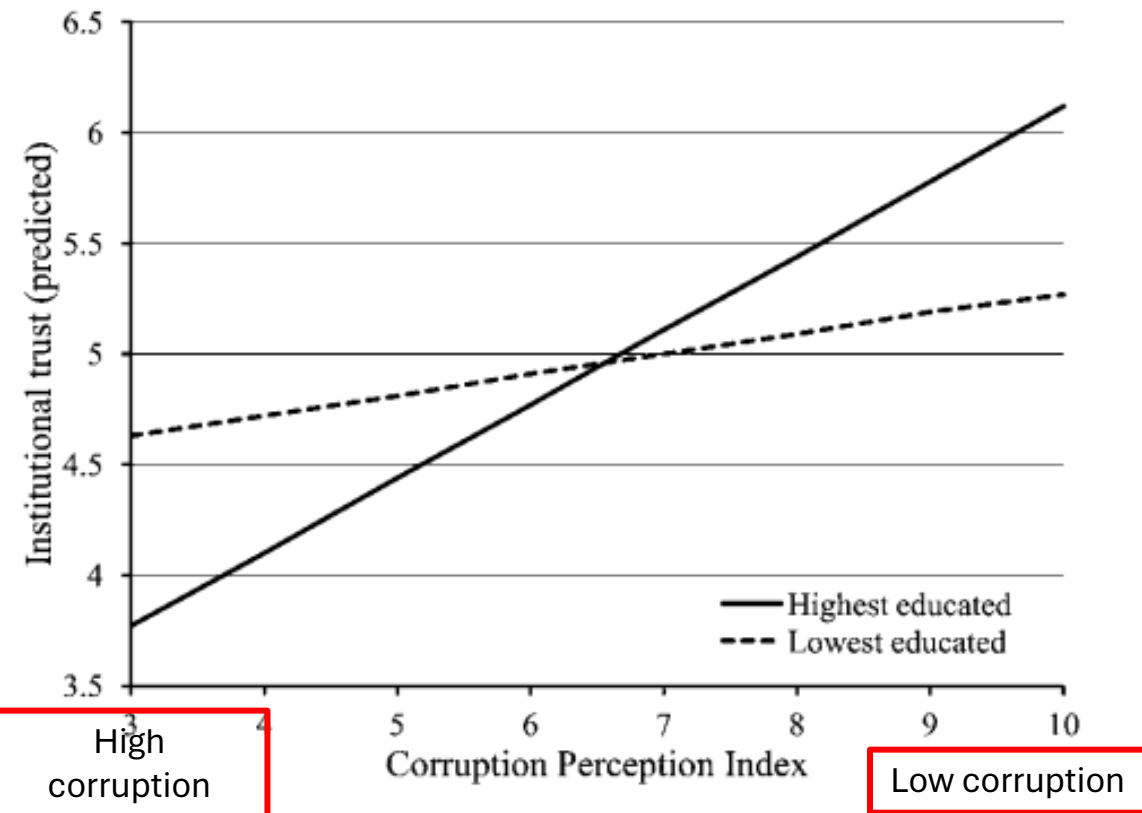
Corruptie

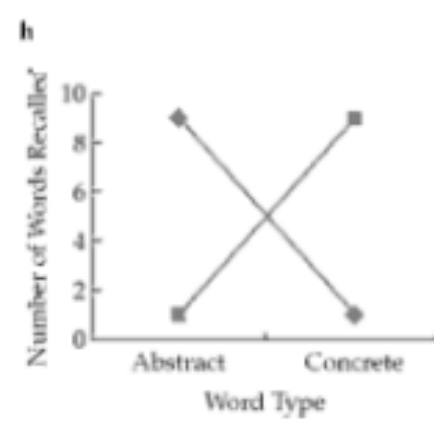
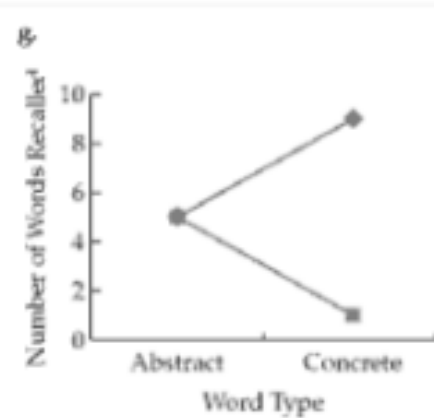
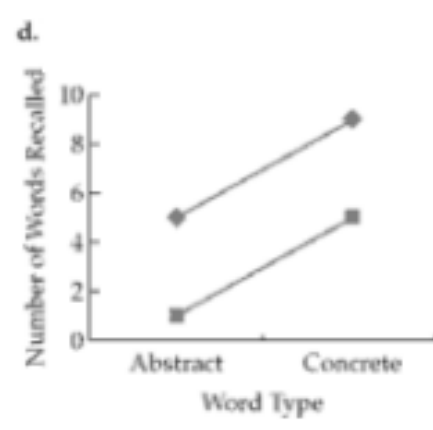
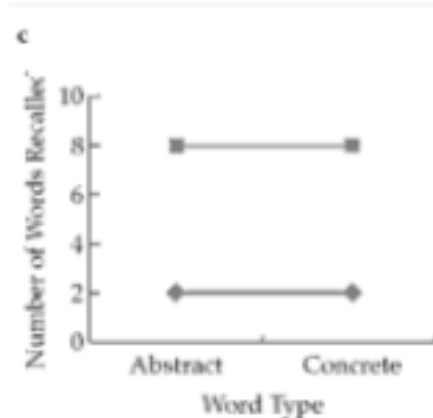
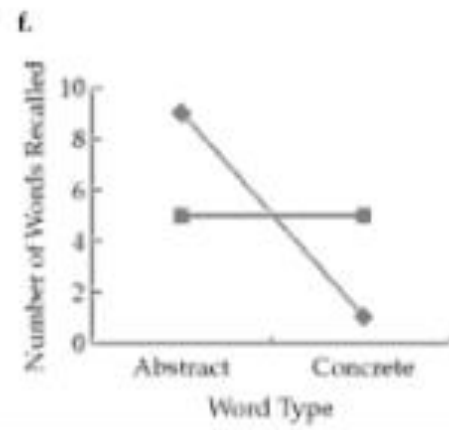
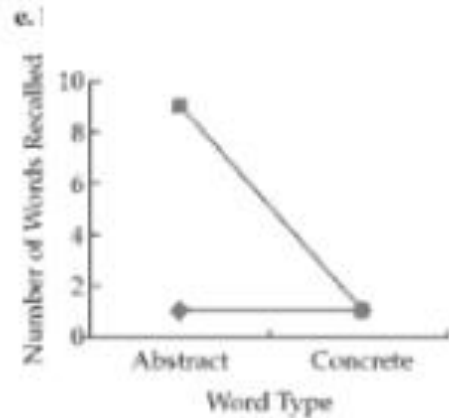
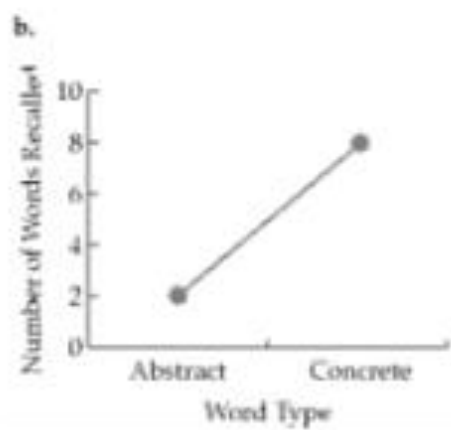
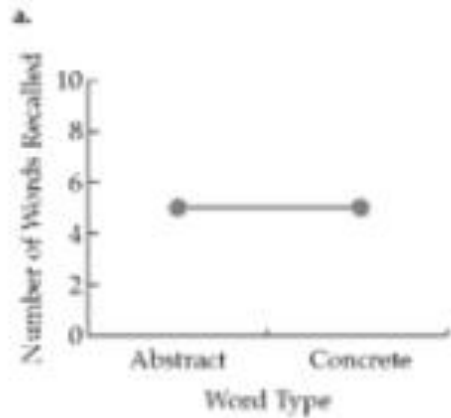
VERTROUWEN



Opleiding

FIGURE 3 Predicted Values of Institutional Trust across Levels of Corruption and Education Groups





INTERACTIES IN REGRESSIE

De regressie-vergelijking



INTERACTIE-EFFECT IN REGRESSIE

Voeg ook altijd het
“main”-effect van
beide variabelen toe!

De interactie is de
vermenigvuldiging tussen
beide variabelen

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$$

Als je herschikt, zie je dat het effect van X1 nu afhankelijk is van X2

$$Y_i = (\beta_0 + \beta_2 X_{2i}) + (\beta_1 + \beta_3 X_{2i}) X_{1i} + \varepsilon_i$$

Effect van X1 wordt nu zelf
beïnvloed door X2

INTERPRETATIE INTERACTIE

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$$

$$Y_i = (\beta_0 + \beta_2 X_{2i}) + (\beta_1 + \beta_3 X_{2i}) X_{1i} + \varepsilon_i$$

$$Y_i = (\beta_0 + \beta_1 X_{1i}) + (\beta_2 + \beta_3 X_{1i}) X_{2i} + \varepsilon_i$$

β_0 : Waarde Y_i als X_{1i} en X_{2i} 0 zijn

β_1 : Effect van X_{1i} als X_{2i} 0 is (dus: hoeveel stijgt Y_i als X_{1i} met 1 omhoog gaat als X_{2i} 0 is)

β_2 : Effect van X_{2i} als X_{1i} 0 is (dus: hoeveel stijgt Y_i als X_{2i} met 1 omhoog gaat als X_{1i} 0 is)

β_3 : Effect van X_{2i} op effect van X_{1i} of Effect van X_{1i} op effect van X_{2i}

(dus met hoeveel meer stijgt Y_i voor iedere toename van X_{1i} met 1 wanneer X_{2i} met 1 omhoog gaat)
of

(dus met hoeveel meer stijgt Y_i voor iedere toename van X_{2i} met 1 wanneer X_{1i} met 1 omhoog gaat)



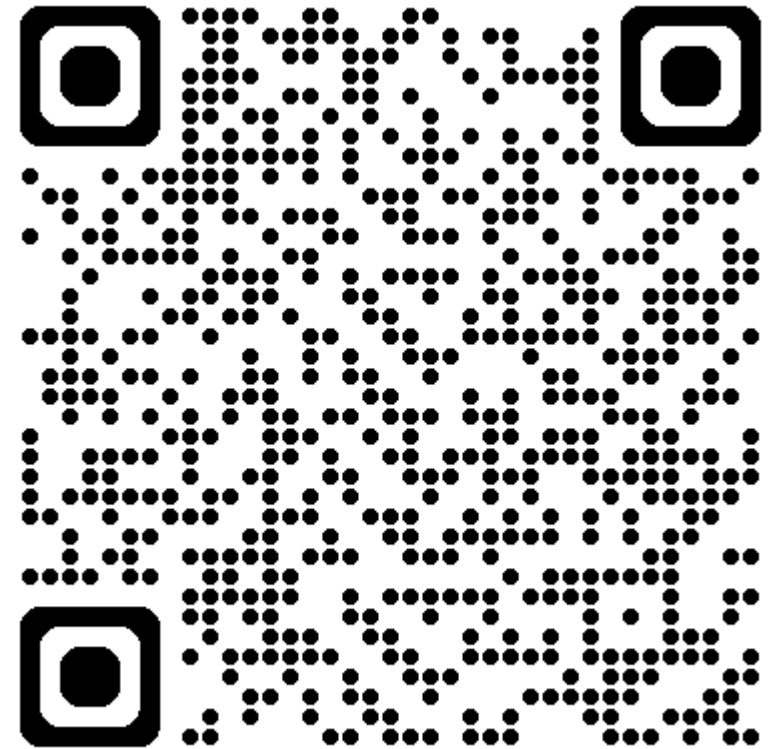
INTERACTIE MET DUMMY

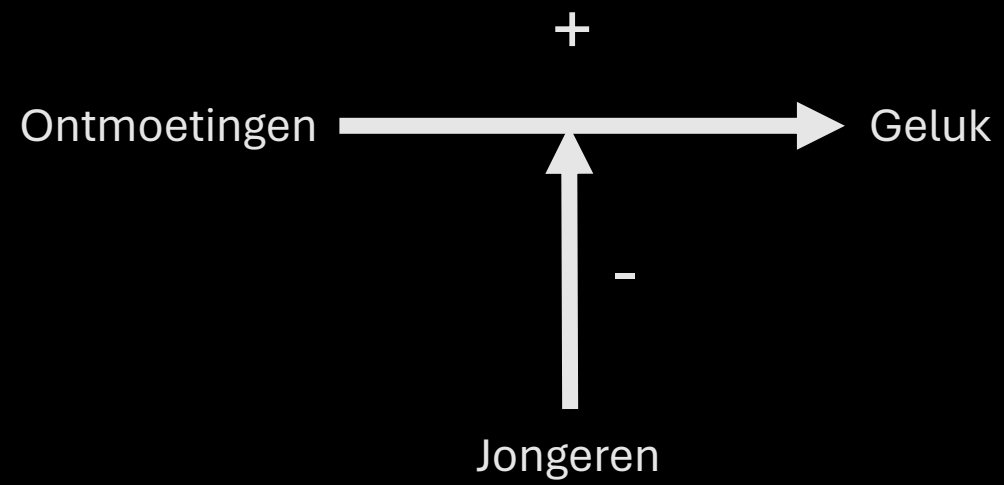
Radical Right, immigration and motivation to control prejudice

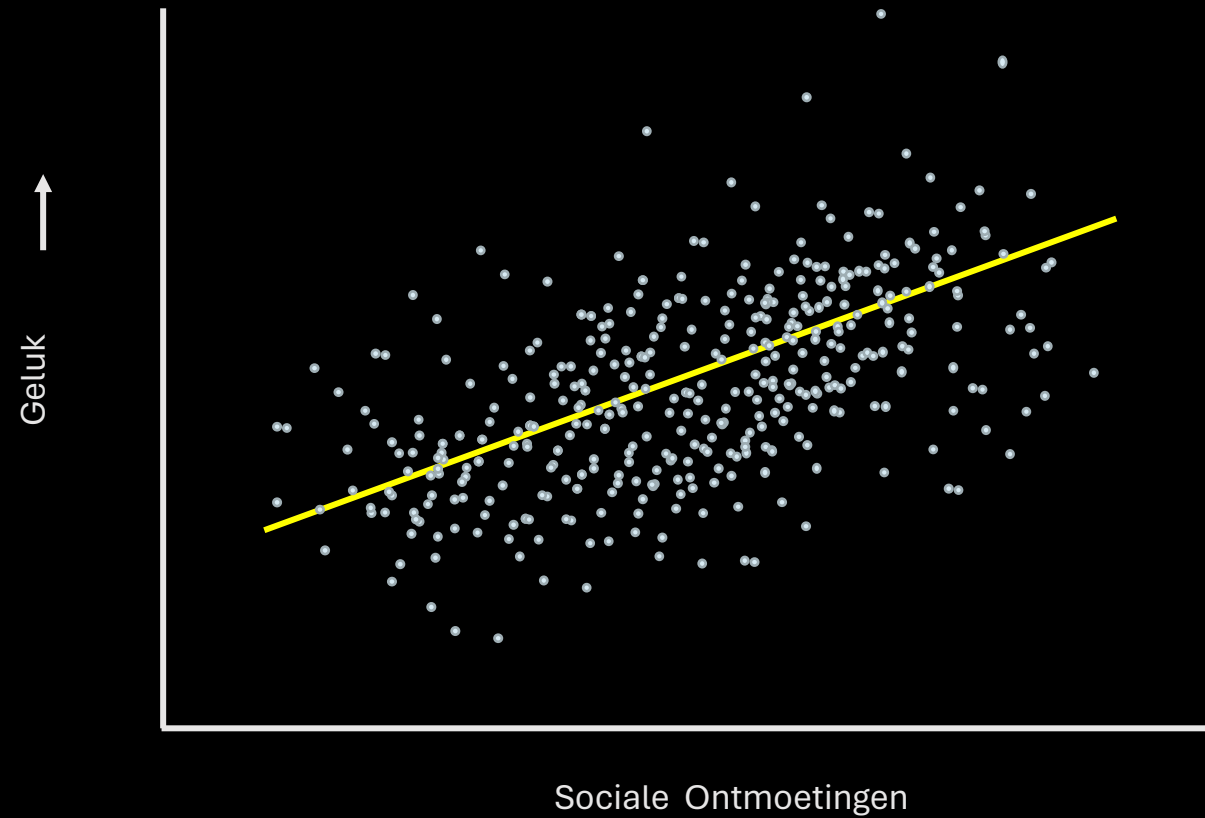
<https://elmarjansen.nl/os>

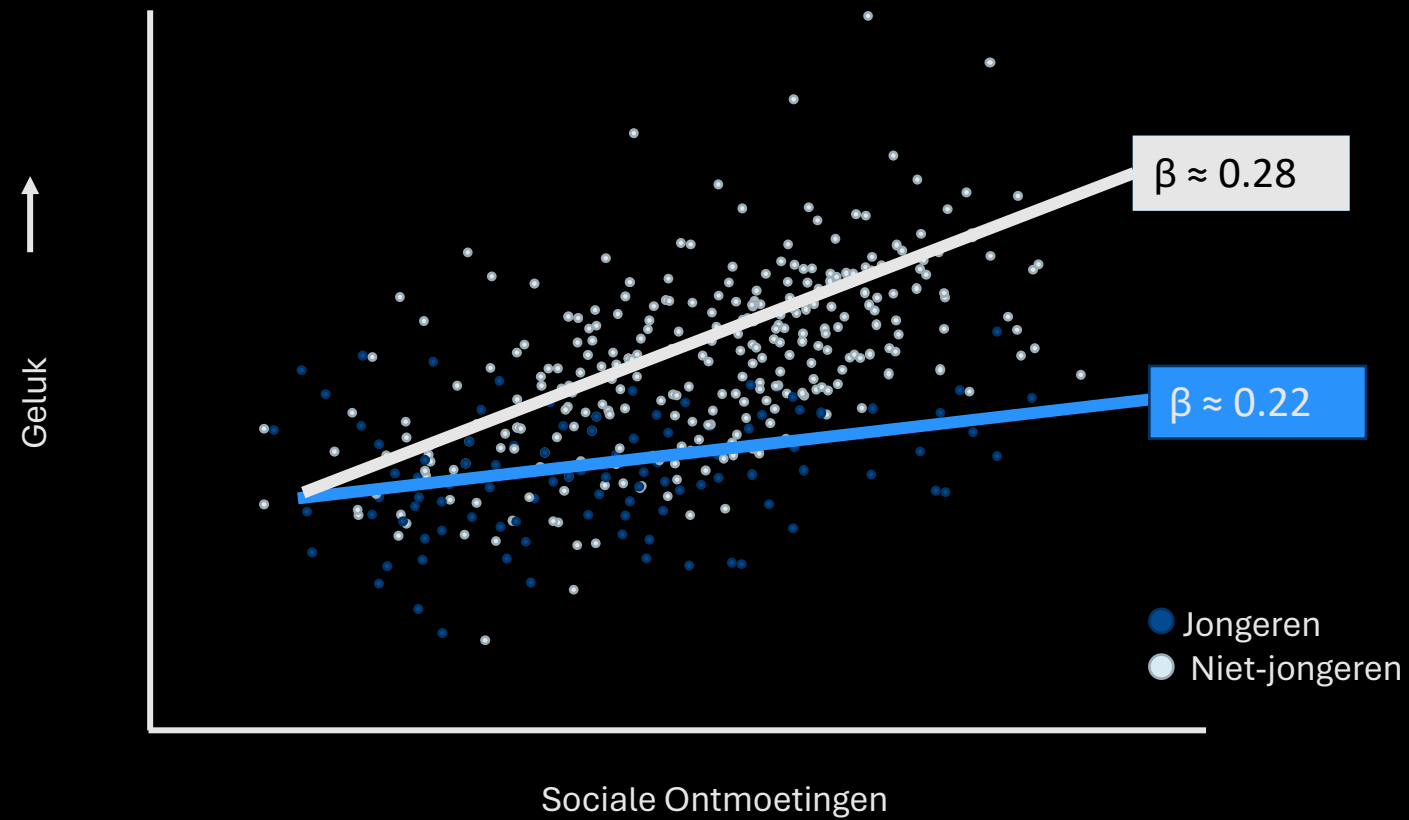
OEFENING 3

Geluk, ontmoetingen en jongeren

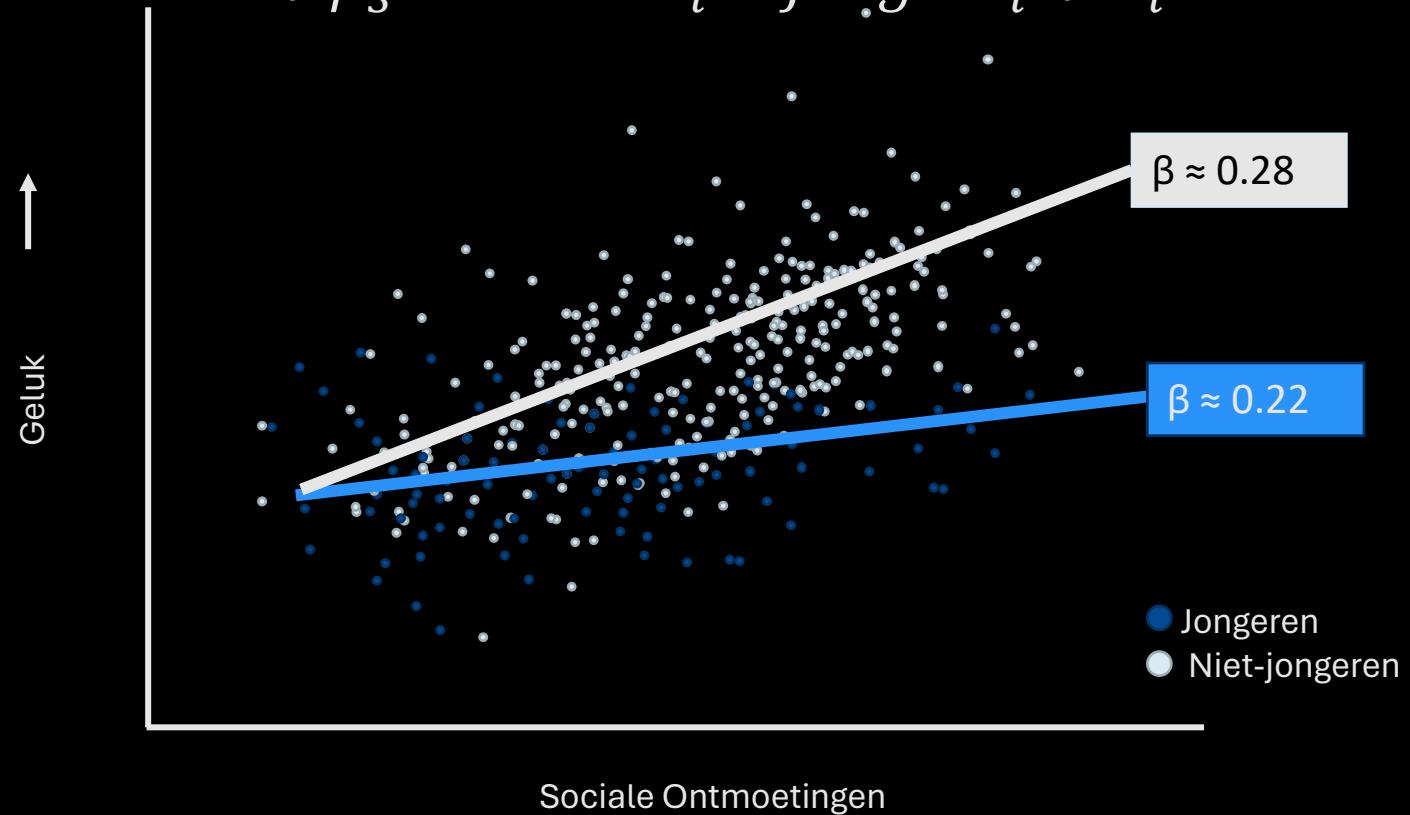




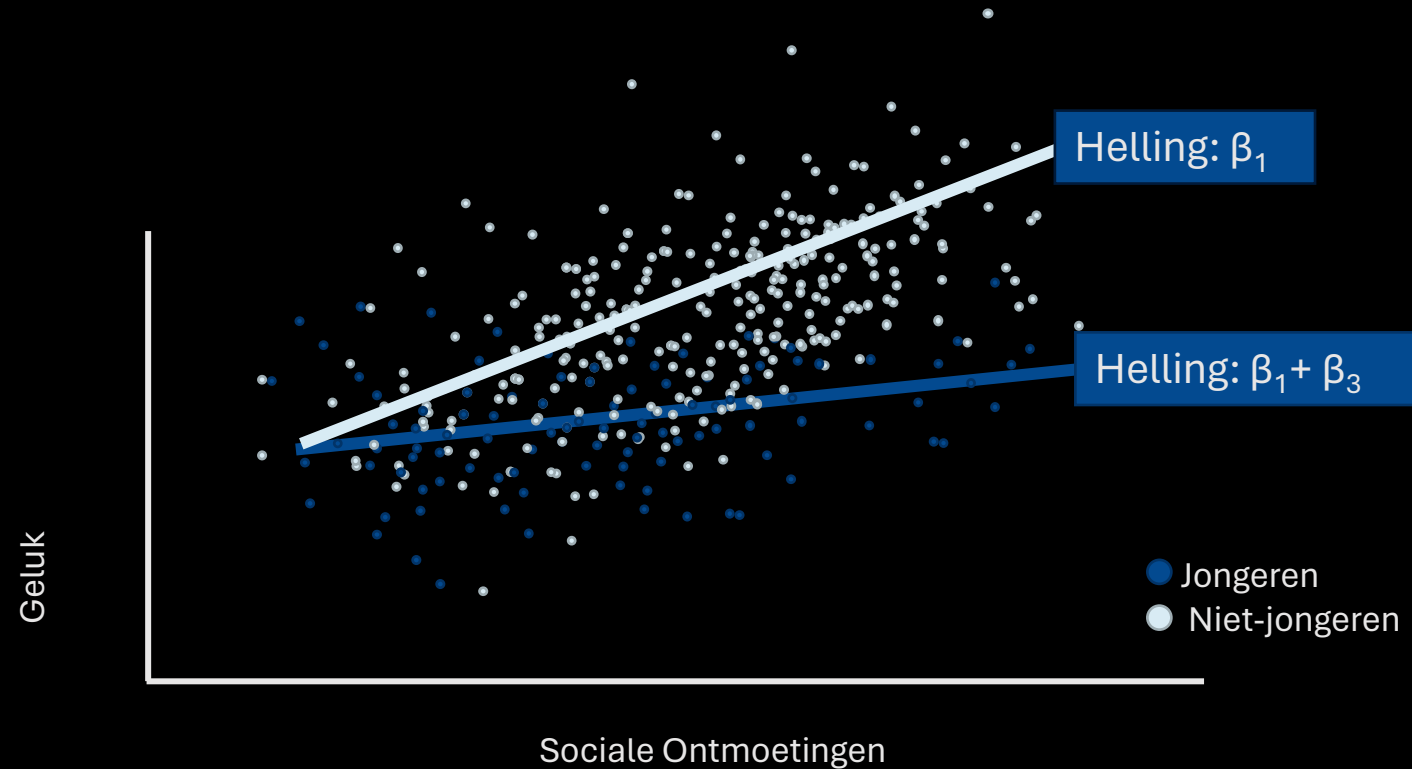




$$Geluk_i = \beta_0 + \beta_1 \times sclmeet_i + \beta_2 \times jongere_i + \beta_3 \times sclmeet_i \times jongere_i + \varepsilon_i$$



$$Geluk_i = \beta_0 + \beta_1 \times sclmeet_i + \beta_2 \times jongere_i + \beta_3 \times sclmeet_i \times jongere_i + \varepsilon_i$$



'Main effect' ontmoetingen:

Effect van **ontmoetingen** als **jongere** 0 is.

'Main effect' jongere:

Effect van **jongere** als **ontmoetingen** 0 is.

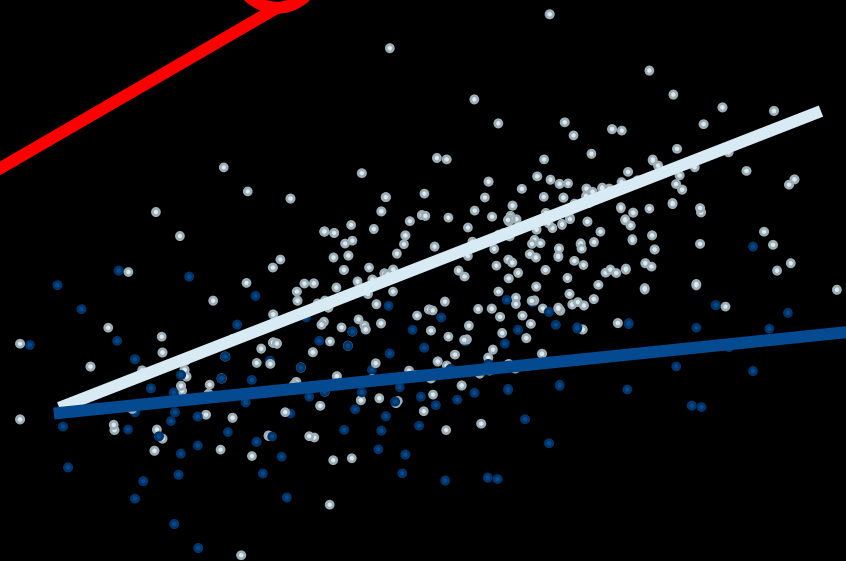
$$Geluk_i = \beta_0 + \beta_1 \times sclmeet_i + \beta_2 \times jongere_i + \beta_3 \times sclmeet_i \times jongere_i + \varepsilon_i$$

'Interactie effect' jongere en ontmoetingen:

Toename effect **ontmoetingen** voor iedere keer dat **jongere** met 1 stijgt

en

Toename effect **jongere** voor iedere keer dat **ontmoetingen** met 1 stijgt.



● Jongeren
● Niet-jongeren

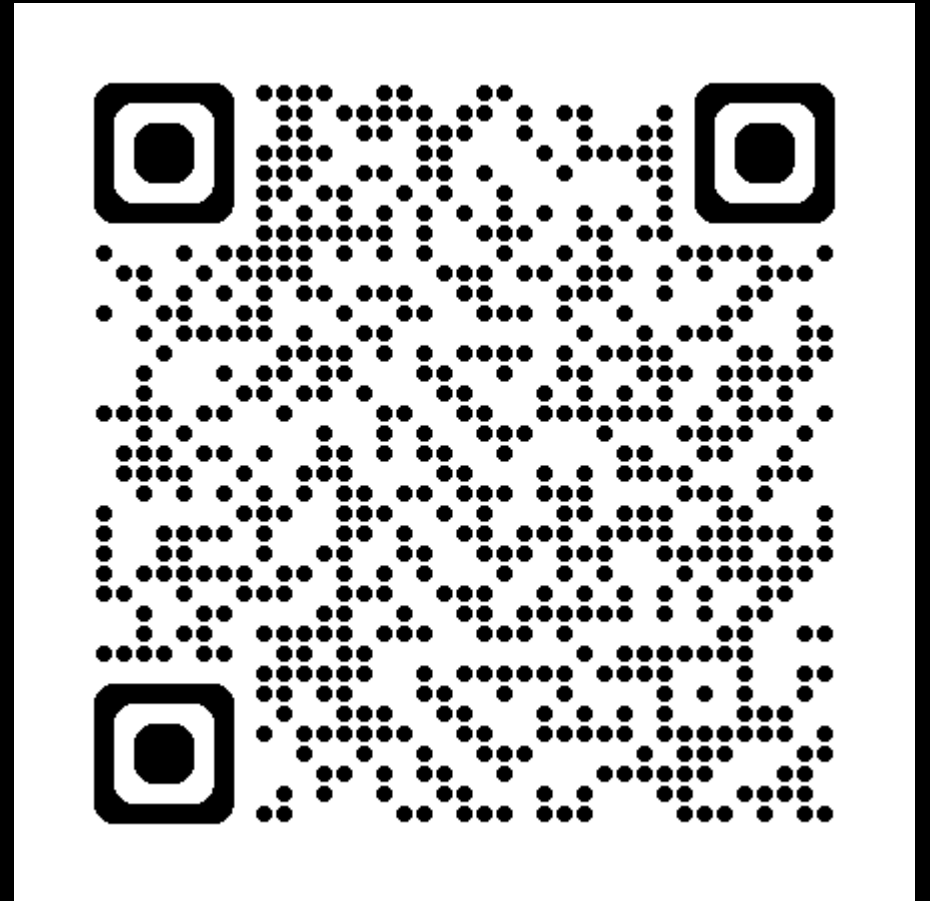
INTERACTIE MET CONTINUE VARIABELE



<https://elmarjansen.nl/os>

OEFENING 4

Geluk, ontmoetingen en leeftijd



Ontmoetingen → Geluk



-

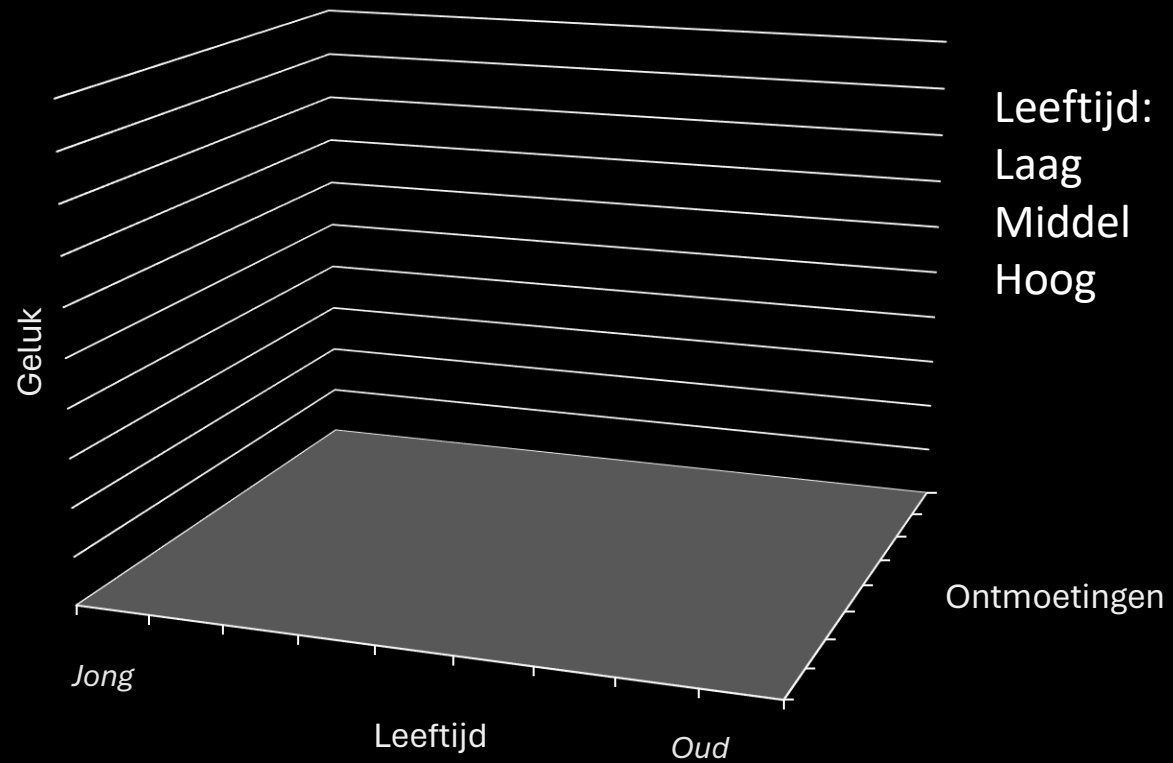
Leeftijd



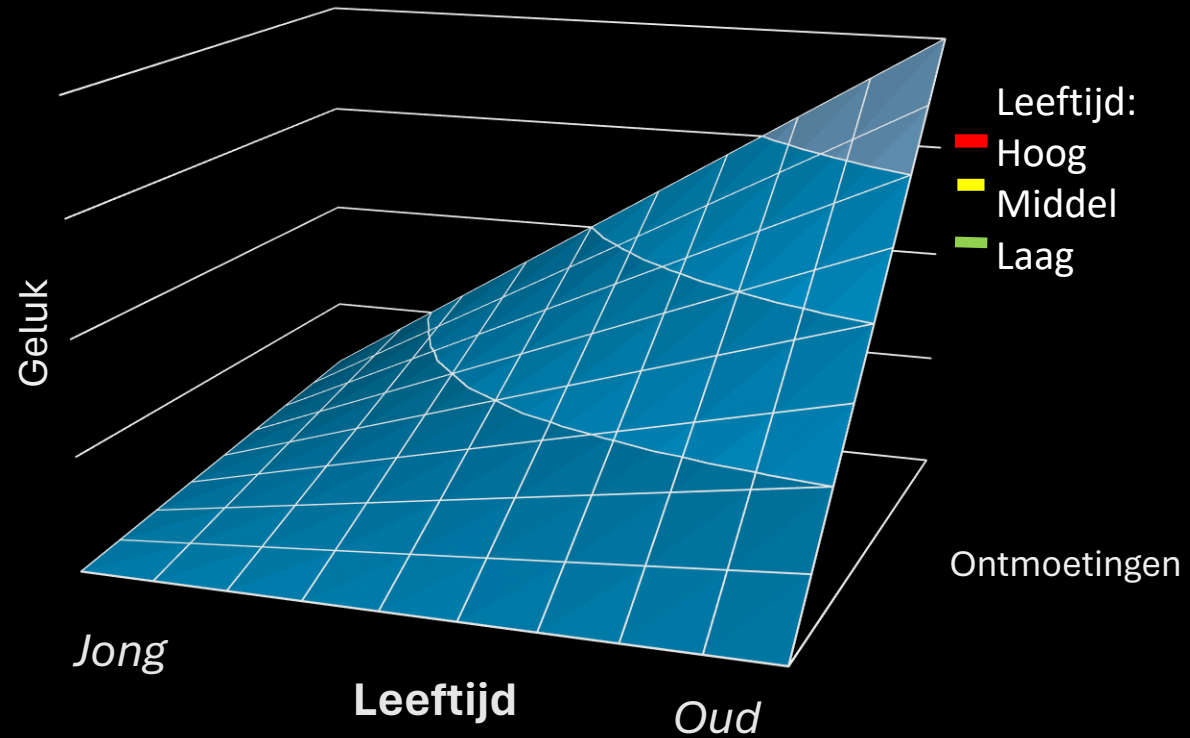
Laag

Hoog

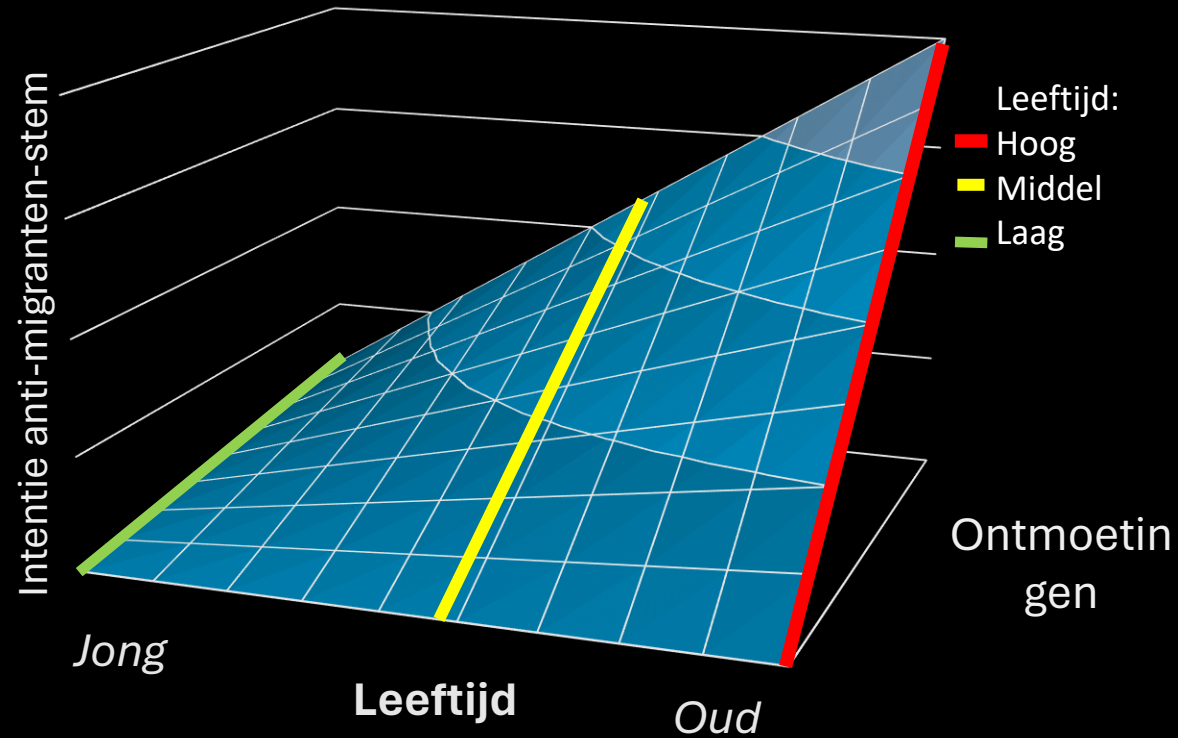
$$Geluk_i = \beta_0 + \beta_1 \times sclmeet_i + \beta_2 \times leeftijd_i + \beta_3 \times sclmeet_i \times leeftijd_i + \varepsilon_i$$



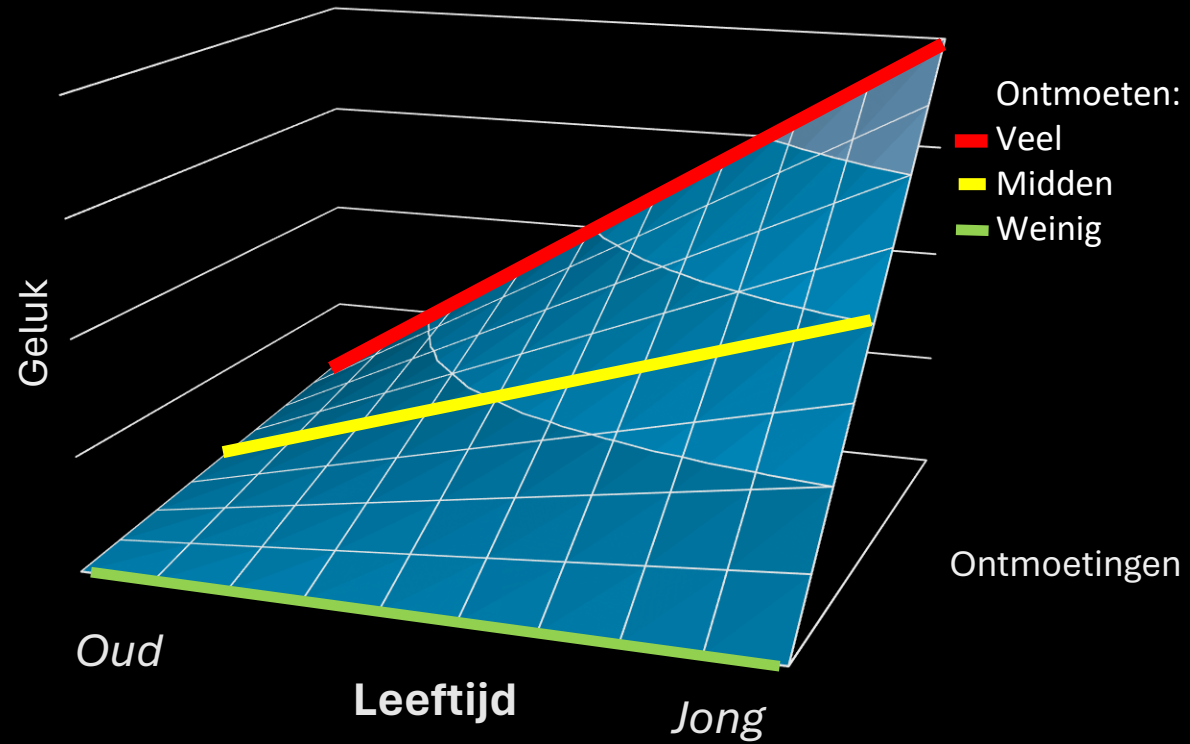
$$Geluk_i = \beta_0 + \beta_1 \times sclmeet_i + \beta_2 \times leeftijd_i + \beta_3 \times sclmeet_i \times leeftijd_i + \varepsilon_i$$



$$Geluk_i = \beta_0 + \beta_1 \times sclmeet_i + \beta_2 \times leeftijd_i + \beta_3 \times sclmeet_i \times leeftijd_i + \varepsilon_i$$



$$VoteRad_i = \beta_0 + \beta_1 Imm_i + \beta_2 MCP_i + \beta_3 Imm_i \times MCP_i + \varepsilon_i$$



INTERACTIE INTERPRETEREN



```
summary(lm(voteSD ~ immscale * mcp, data=dt))
```

Call:

```
lm(formula = voteSD ~ immscale * mcp, data = dt)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-9.3863	-0.8441	-0.0837	0.1008	10.0587

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1.08813	0.53371	-2.039	0.04166	*
immscale	3.45892	0.20787	16.640	< 2e-16	***
mcp	0.41402	0.12701	3.260	0.00114	**
immscale:mcp	-0.69382	0.05489	-12.640	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.152 on 1377 degrees of freedom
(9 observations deleted due to missingness)

Multiple R-squared: 0.3973, Adjusted R-squared: 0.396

F-statistic: 302.6 on 3 and 1377 DF, p-value: < 2.2e-16

$$\begin{aligned}
 \text{VoteRad}_i = & \\
 & \beta_0 + \\
 & \beta_1 \times \text{Imm}_i + \\
 & \beta_2 \times \text{MCP}_i + \\
 & \beta_3 \times \text{Imm}_i \times \text{MCP}_i + \varepsilon_i
 \end{aligned}$$

Effect van anti-immigratie-mening wordt significant gemodereerd door MCP
 of
 Effect van MCP is significant gemodereerd door anti-immigratie houding

$$-1.1 + 3.5 \text{Imm}_i + 0.4 \text{MCP}_i - 0.7 \text{Imm}_i \text{MCP}_i + \varepsilon_i$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1.08813	0.53371	-2.039	0.04166	*
immscale	3.45392	0.20787	16.640	< 2e-16	***
mcp	0.41402	0.12701	3.260	0.00114	**
immscale:mcp	-0.69382	0.05489	-12.640	< 2e-16	***

INTERPRETATIE INTERACTIE-EFFECT

In het algemeen:

Het **interactie-effect** staat voor de verandering in het effect van X_1 op Y bij iedere toename van één eenheid in X_2

Met iedere eenheid toename in MCP, daalt het effect van anti-immigrant houding met 0.7.

=

$$-1.1 + 3.5Imm_i + 0.4MCP_i - 0.7Imm_iMCP_i + \varepsilon_i$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1.08813	0.53371	-2.039	0.04166	*
immscale	3.45892	0.20787	16.640	< 2e-16	***
mcp	0.41402	0.12701	3.260	0.00114	**
immscale:mcp	-0.69382	0.05489	-12.640	< 2e-16	***

INTERPRETATIE MAIN-EFFECTS

In het algemeen:

Het **main effect** van een variabele in de interactie is het effect van die variabele,
als de andere variabele in de interactie 0 is.

$$= -1.1 + 3.5Imm_i + 0.4MCP_i - 0.7Imm_iMCP_i + \varepsilon_i$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1.08813	0.53371	-2.039	0.04166	*
immscale	3.45892	0.20787	16.640	< 2e-16	***
mcp	0.41402	0.12701	3.260	0.00114	**
immscale:mcp	-0.69382	0.05489	-12.640	< 2e-16	***

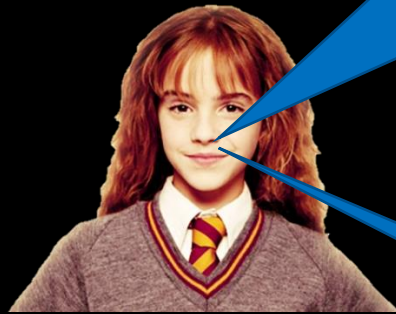
Main effect anti-immigration-houding:
Effect van anti-immigratie-houding,
als MCP 0 is.

Main effect MCP:
Effect van MCP, als anti-
immigratie-houding 0 is.

$$VoteRad_i = -1.1 + 3.5Imm_i + 0.4MCP_i - 0.7Imm_iMCP_i + \varepsilon_i$$

Dit is waarom main effecten
meestal niet zo interessant zijn.

Door de interactie zijn ze nu
'voorwaardelijk', het main-effect
is alleen het effect van die
variabele onder een heel
specifieke voorwaarde, namelijk
dat hij 0 is.



Vaak kan dat
niet eens...

Als MCP = 1:

$$VoteRad_i =$$

$$-1.1 + 3.5Imm_i + 0.4 \times 1 - 0.7Imm_i \times 1 + \varepsilon_i$$

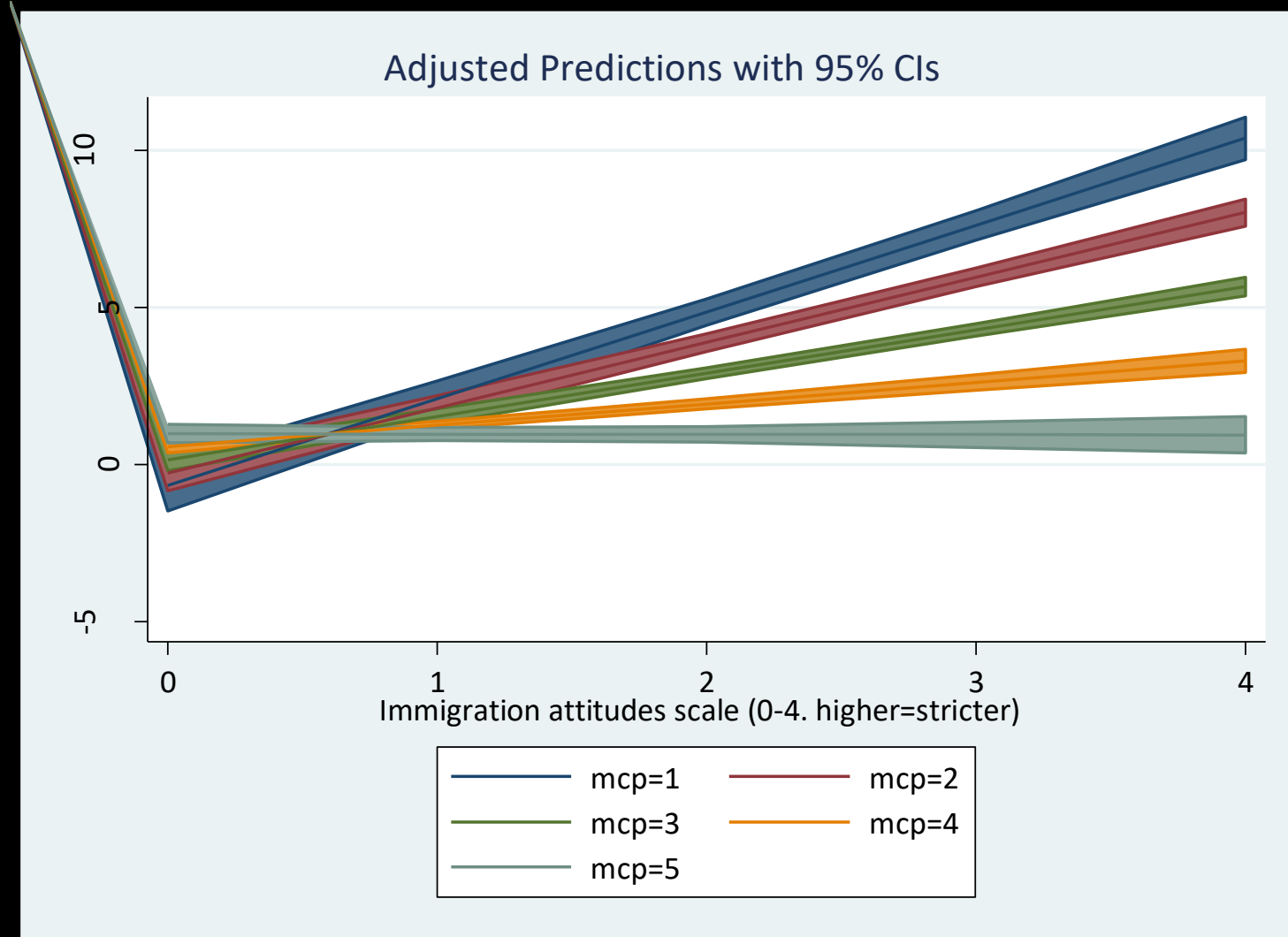
$$= -0.7 + 2.8Imm_i + \varepsilon_i$$

Als MCP = 5:

$$VoteRad_i =$$

$$-1.1 + 3.5Imm_i + 0.4 \times 5 - 0.7Imm_i \times 5 + \varepsilon_i$$

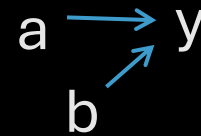
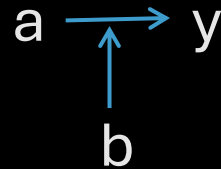
$$= 0.9 - 0.01Imm_i + \varepsilon_i$$



CHECK-LIST INTERACTIES

- ✓ Heb je **theoretische redenen** om een echte *interactie* te verwachten? (of gaat het “gewoon” om een additioneel effect?)

(Dus dit: niet dit:)



- ✓ Laat altijd je twee “**main-effects**” ook meedoen in de analyse
- ✓ **Interpretatie van main-effects** is meestal niet betekenisvol: ze hangen af van het andere effect.

LINEAIRE REGRESSIE: DUMMIES EN INTERACTIES

30 mei 2024

Training O + S

Elmar Jansen (elmar@elmarjansen.nl)